SECOND EDITION

Handbook of Laser Technology and Applications

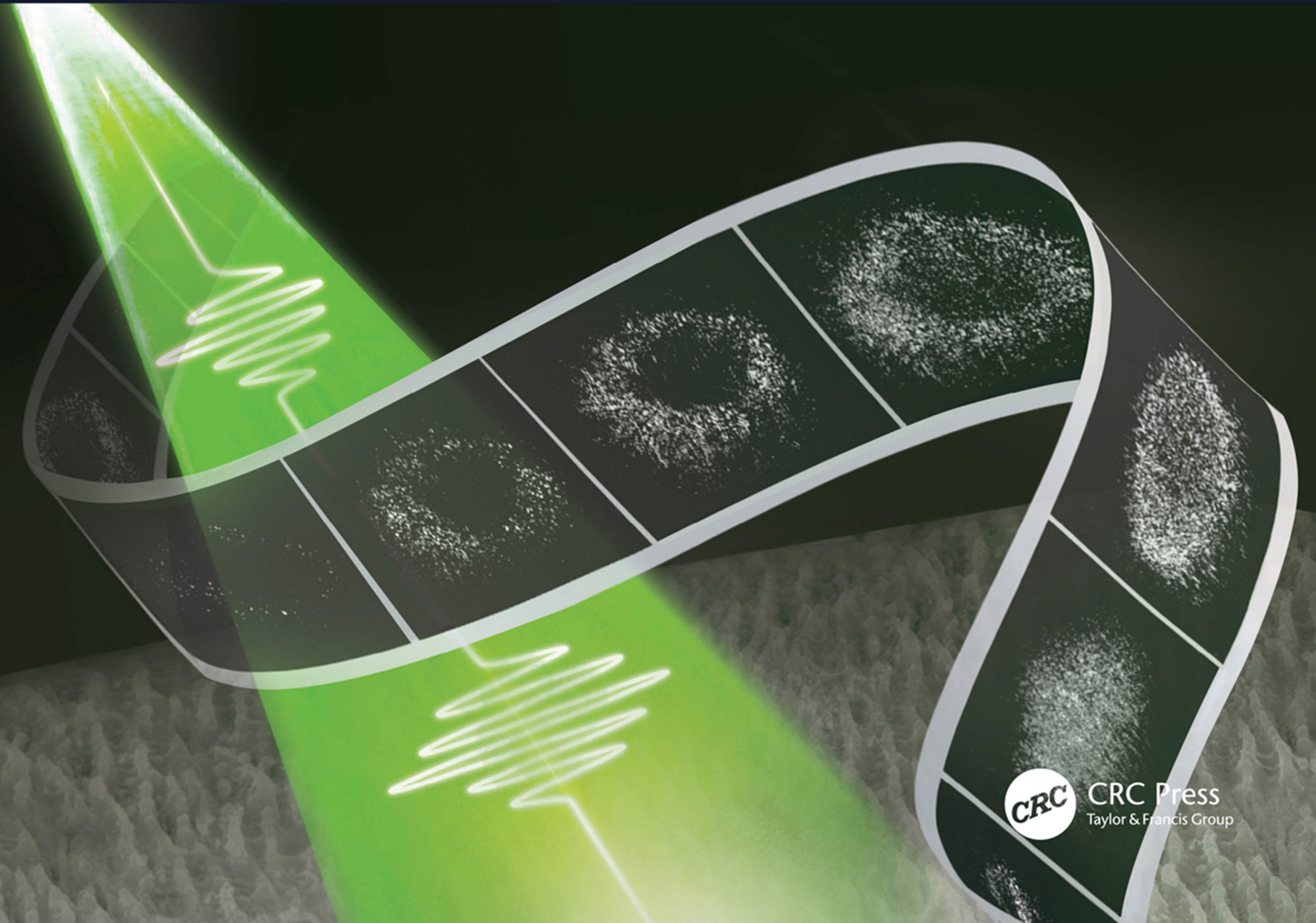
VOLUME IV

Laser Applications

Medical, Metrology and Communication

edited by

Chunlei Guo • Subhash Chandra Singh



CRC Press
Taylor & Francis Group

Handbook of Laser Technology and Applications

Handbook of Laser Technology and Applications

Lasers: Principles and Operations (Volume One)
Second Edition

Lasers Design and Laser Systems (Volume Two)
Second Edition

Laser Applications: Material Processing and Spectroscopy (Volume Three)
Second Edition

Laser Applications: Medical, Metrology and Communication (Volume Four)
Second Edition

Handbook of Laser Technology and Applications

Laser Applications: Medical, Metrology and Communication (Volume Four)

Second Edition

Edited by

Chunlei Guo

Subhash Chandra Singh

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Second edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

© 2021 Taylor & Francis Group, LLC

First edition published by IOP Publishing 2004

CRC Press is an imprint of Taylor & Francis Group, LLC

Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, access www.copyright.com or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. For works that are not available on CCC please contact mpkbookspermissions@tandf.co.uk

Trademark notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data
Names: Guo, Chunlei, editor. | Singh, Subhash Chandra, editor.
Title: Handbook of laser technology and applications : four volume set /
[edited by] Chunlei Guo and Subhash Chandra Singh.
Description: 2nd edition. | Boca Raton : CRC Press, 2021- |
Series: Handbook of laser technology and applications | Includes bibliographical
references and index. | Contents: v. 1. Lasers: principles and
operations—v. 2. Laser design and laser systems—v. 3. Lasers
applications: materials processing—v. 4. Laser applications: medical, metrology a [?].
Identifiers: LCCN 2020037189 (print) | LCCN 2020037190 (ebook) |
ISBN 9781138032613 (v. 1 ; hardback) | ISBN 9781138032620 (v. 2 ; hardback) |
ISBN 9781138033320 (v. 3 ; hardback) | ISBN 9780367649173 (v. 4 ; hardback) |
ISBN 9781138196575 (hardback) | ISBN 9781315389561 (v. 1 ; ebook) |
ISBN 9781003127130 (v. 2 ; ebook) | ISBN 9781315310855 (v. 3 ; ebook) |
ISBN 9781003130123 (v. 4 ; ebook)
Subjects: LCSH: Lasers.
Classification: LCC TK7871.3 .H25 2021 (print) | LCC TK7871.3 (ebook) |
DDC 621.36/6—dc23
LC record available at https://lccn.loc.gov/2020037189
LC ebook record available at https://lccn.loc.gov/2020037190

ISBN: 9780367649173 (hbk)
ISBN: 9780367655631 (pbk)
ISBN: 9781003130123 (ebk)

Typeset in Times
by codeMantra

Contents

Preface vii
Editors ix
Contributors xi

1. **Lasers in Metrology: Section Introduction** 1
Julian Jones and Lance Thomas

2. **Fundamental Length Metrology** 3
Jens Flügge, Stefanie Kroker, and Harald Schnatz

3. **Laser Velocimetry** 23
Cameron Tropea

4. **Laser Vibrometers** 43
Neil A. Halliwell

5. **Electronic Speckle Pattern Interferometry (ESPI)** 61
Lianxiang Yang and Xinya Gao

6. **Optical Fibre Hydrophones** 81
Geoffrey A. Cranch and Philip J. Nash

7. **Laser Stabilization for Precision Measurements** 111
G. P. Barwood and P. Gill

8. **Laser Cooling and Trapping** 127
Charles Adams and Ifan Hughes

9. **Precision Timekeeping: Optical Atomic Clocks** 139
Shimon Kolkowitz and Jun Ye

10. **Optical Atomic Clock and Laser Applications to Length and Time Metrology** 157
Y. Jiang, Y. Huang, and K. Gao

11. **Gravitation Measurements with Laser Interferometry** 171
Harold V. Parks

12. **Satellite Laser Ranging** 181
José C. Rodríguez and Graham M. Appleby

13. **Lasers in Medical: Section Introduction** 199
Terence A. King and Brian C. Wilson

14. **Light–Tissue Interactions** 201
Steven Jacques and Michael Patterson

15. **Ophthalmic Laser Therapy and Surgery** 227
Daniel Palanker

16. **Therapeutic Application: Refractive Surgery** 243
Leonardo Mastropasqua

17. Photodynamic Therapy 249
Brian C. Wilson and Stephen G. Bown

18. Therapeutic Applications: Thermal Treatment of Tumours 261
Stephen G. Bown

19. Therapeutic Applications: Dermatology—Selective Photothermolysis 267
Sean Lanigan

20. Therapeutic Applications: Lasers in Vascular Surgery 273
Mahesh Pai

21. Therapeutic Applications: Free-Electron Laser 281
E. Duco Jansen, Michael Copeland, Glenn S. Edwards, William Gabella, Karen Joos, Mark A. Mackanos, Jin H. Shen, and Stephen R. Uhlhorn

22. Medical Diagnostics 291
Brian C. Wilson

23. Broad Bandwidth Light Sources in Optical Coherence Tomography (OCT) 309
Jinxin Huang and Jannick P. Rolland

24. Laser Applications in Biology and Biotechnology 321
Sebastian Wachsmann-Hogiu, Alexander J Annala, and Daniel L Farkas

25. Biomedical Laser Safety 345
Harry Moseley and Bill Davies

26. Laser in Communications: Section Introduction 365
John Marsh, Subhash C. Singh, and Chunlei Guo

27. Fibre-to-the-Chip: Development of Vertical Cavity Surface-Emitting Laser Arrays Designed for Integration with VLSI Circuits 367
Ashok Krishnamoorthy, L. M. F. Chirovsky, K. W. Goosen, J. Lopata, and W. S. Hobson

28. Advances in Laser Satellite Communications 383
Hamid Hemmati

29. Passive Silicon Photonic Integrated Components and Circuits for Optical Communications 397
Daoxin Dai, Yiwei Xie, and Yaocheng Shi

30. Fibre-Optic Transmission Systems from Chip-to-Chip Interconnects to Trans-Oceanic Cables 427
Peter J. Winzer

31. Visible Light Communications and LiFi 443
Cheng Chen, Mohamed Sufyan Islim, and Harald Haas

32. Optical Data Storage 463
Byoung S. Ham

Index 473

Preface

This updated Handbook comes at the time when the world just celebrated the 60th anniversary of the laser. Compared to most fields in science and technology, the laser is still a relatively young one, but its developments have been astonishing. Today, hardly any area of modern life is left untouched by lasers, so it is almost impossible to provide a complete account of this subject.

As challenging as it is, this updated Handbook attempts to provide a comprehensive coverage on modern laser technology and applications, including recent advancements and state-of-the-art research and developments. The main goal of developing this Handbook is to provide both an overview and details of ever-expanding technologies and applications in lasers.

We want this Handbook to be useful for both newcomers and experts in lasers. To meet these goals, the chapters in this Handbook are typically developed in a style that does not require advanced mathematical tools. On the other hand, they are written by the experts in each area so that the most important concepts and developments are covered.

The first edition of the Handbook was released in 2003. It has been hugely popular and ranked as one of the top ten most referenced materials by the publisher. Eighteen years later, although a relatively short period for many more established scientific fields, the Handbook has become outdated, and an update is overdue. The rapid changes in lasers are certainly reinforced by my own experience of teaching and researching the subject in the Institute of Optics at University of Rochester. Flipping through my old lecture notes on lasers, I am often amazed at how much progress we have witnessed in this field over the years.

I am indebted to the editors of the first edition, Colin Webb and Julian Jones, who brought this original Handbook into existence. When I was asked to take over this second edition, it laid before me a daunting task of how to rejuvenate the Handbook while keeping its original flavour. Since many of the fundamental principles of the laser are well established, we tried to honour the original authors by keeping the chapters on fundamental concepts where possible. If a revision is needed, we usually started by asking the original authors for the revision but if impossible, we brought in new authors to revise these chapters.

As the laser shines in modern applications, we added a large number of new chapters reflecting the most recent advancements in laser technologies. Throughout the Handbook, entirely new sections were added, including sections on materials processing, laser spectroscopy and lasers in imaging and communications. Nearly all chapters in these sections are either entirely new or substantially revised. On the other hand, some of the topics previously included have seen dwindling relevance today. We had to make the hard decision to let go of some of these outdated chapters from the first edition. Despite these deletions, this new Handbook still grows significantly from the original three volumes to the current four volumes.

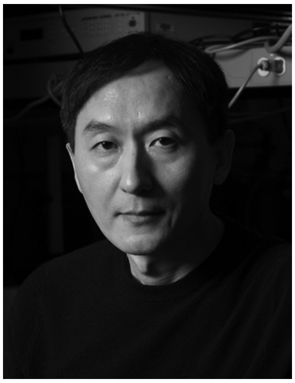
Bringing this large project to its conclusion is the collective efforts of many individuals. It began with the encouragement and guidance of Lu Han, the then managing editor of this Handbook. I know how much Lu cared about this project. I still remember an initial phone call with Lu, we finished it at a late afternoon past 5 pm. Over the phone, I was told that I would receive the first edition of this Handbook. To my surprise, I had the handbooks in my hand the next morning. At CRC press, this project was later passed onto Carolina Antunes and finally to Lara Spieker, who has been essential in bringing this project to its conclusion.

Many people have provided me with indispensable help. My co-editor, Subhash C. Singh, at the University of Rochester, helped chart the layout of this new edition and worked along with me throughout this project. Ying Zhang, who was a senior editor at Changchun Institute of Optics, Fine Mechanics, and Physics (CIOMP) in China, spent a half year with us in Rochester, where his years of professional editorial experience helped move this project forward significantly. Lastly, my thanks go to Pavel Redkin of CIOMP, who made significant contributions in communicating with the chapter authors and guiding them throughout the project. Additionally, my appreciation goes to Kai Davies, Sandeep K. Maurya, Xin Wei, and Wenting Sun for their help in this Handbook project.

Chunlei Guo
Editor-in-Chief
University of Rochester

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Editors



Chunlei Guo is a Professor in The Institute of Optics and Physics at the University of Rochester. Before joining the Rochester faculty in 2001, he earned a PhD in Physics from the University of Connecticut and did his postdoctoral training at Los Alamos National Laboratory. His research is in studying femtosecond laser interactions with matter, spanning from atoms and molecules to solid materials.

His research at University of Rochester has led to the discoveries of a range of highly functionalized materials through femtosecond laser processing, including the so-called black and colored metals and superhydrophillic and superhydrophobic surfaces. These innovations may find a broad range of applications, and have also been extensively featured by the media, including multiple New York Times articles. Lately, he devoted a significant amount of efforts to developing technologies for global sanitation by working with the Bill & Melinda Gates Foundation. Through this mission, he visited Africa multiple times to understand humanitarian issues. To further expand global collaboration under the Gates project, he helped establish an international laboratory at Changchun Institute of Optics, Fine Mechanics, and Physics in China. He is a Fellow of the American Physical Society, Optical Society of America, and International Academy of Photonics & Laser Engineering. He has authored about 300 referred journal articles..

Subhash Chandra Singh is a scientist at the Institute of Optics, University of Rochester and an Associate Professor at Changchun Institute of Optics, Fine Mechanics, and Physics. Dr. Singh earned a Ph.D. in Physics from University of Allahabad, India in 2009. Prior to working with the Guo Lab, he was IRCSET-EMPOWER Postdoctoral Research Fellow at Dublin City University, Ireland for 2 years and a DST-SERB Young Scientist at University of Allahabad for 3 years. He has more than 10 years of research experience in the fields of laser-matter interaction, plasma, nanomaterial processing, spectroscopy, energy applications, plasmonics, and photonics. He has published more than 100 research articles in reputable refereed journals and conference proceedings. His past editor experience includes serving as the main editor for Wiley-VCH book *Nanomaterials; Processing and Characterization with Lasers* and guest editor for special issues of a number of journals.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contributors

Charles Adams
Alexander J. Annala
Graham M. Appleby
G. P. Barwood
Stephen G. Bown
Cheng Chen
L. M. F. Chirovsky
Michael Copeland
Geoffrey A. Cranch
Daoxin Dai
Bill Davies
Glenn S. Edwards
Daniel L. Farkas
Jens Flügge
William Gabella
K. Gao
Xinya Gao
P. Gill
K. W. Goosen
Chunlei Guo
Harald Haas
Neil A. Halliwell
Byoung S. Ham
Hamid Hemmati
W. S. Hobson
Jinxin Huang
Y. Huang
Ifan Hughes
Mohamed Sufyan Islim
Steven Jacques
E. Duco Jansen
Y. Jiang
Julian Jones
Karen Joos
Terence A. King
Shimon Kolkowitz
Ashok Krishnamoorthy
Stefanie Kroker
Sean Lanigan
J. Lopata
Mark A. Mackanos
John Marsh
Leonardo Mastropasqua
Harry Moseley
Philip J. Nash
Mahesh Pai
Daniel Palanker
Harold V. Parks
Michael Patterson
José C. Rodríguez
Jannick P. Rolland
Harald Schnatz
Jin H. Shen
Yaocheng Shi
Subhash C. Singh
Lance Thomas
Cameron Tropea
Stephen R. Uhlhorn
Sebastian Wachsmann-Hogiu
Brian C. Wilson
Peter J. Winzer
Yiwei Xie
Lianxiang Yang
Jun Ye

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

Lasers in Metrology: Section Introduction

Julian Jones and Lance Thomas

From the earliest inception of the laser, its extraordinary potential to revolutionize optical measurement techniques was recognized. In one sense, the laser can be regarded as a stable oscillator, generating electromagnetic radiation of stable frequency. It is this characteristic which is the basis of the use of lasers in length metrology, the topic of Chapter 2 by Jens Flügge et al.

Beyond fundamental length metrology, laser interferometry is at the heart of many high-resolution engineering measurements [1,2]. As examples of these, Cam Tropea and Neil Halliwell describe laser velocimetry and laser vibrometry in Chapters 3 and 4. In each case, a laser is used to probe a remote target or measurement volume to determine the motion of a solid surface or fluid. Most of the techniques that they describe are based on interferometry or can be equivalently described in terms of the Doppler frequency shift occurring in radiation scattered from a moving body. In all cases, a considerable advantage of the optical measurement is that it is non-intrusive: photons exert negligible perturbation on the system under measurement—unlike, for example, the mass-loading caused by the attachment of mechanical vibrometres.

Many interferometric techniques have been developed to produce full-field interferograms, mapping out the optical phase over an extended area. Full-field interferometry was first developed for testing the shape of optical surfaces. For targets that are not interferometrically smooth, holographic interferometry may be used for measurement of changes in the optical path length. More conveniently, surface shape and motion can be measured in real time by speckle correlation interferometry, in which a speckle pattern produced by laser light scattered from a test surface mixes with a reference beam to form an image on a camera, which is then correlated with a similar image from the test object in a deformed state, from which the deformation is revealed. This is the subject of electronic speckle interferometry, described by David Towers and Clive Buckberry in Chapter 5.

The advent of optical fibre technology brought about a renaissance in optical instrumentation. Perhaps, the most important application is in hydrophones. Optical fibres are used to form laser interferometers. The effect of changes in hydrostatic pressure is to modulate the relative path length within the fibres and, hence, to produce a phase change. Highly sensitive sensing elements, with controllable geometry, and with the ability to multiplex large numbers of sensors onto a single-fibre downlead, are thus realized, as described in Geoffrey Cranch and Phil Nash's Chapter 6.

In a subject as diverse as laser optical instrumentation, this section can do no more than providing a representative selection. Laser stabilization for precision measurements is given in Chapter 7, while Chapter 8 describes laser cooling and trapping. The cold atoms are the basis of atomic clocks described in Chapter 9.

Today, lasers have enabled time and length standards (Chapter 10). Length is fundamentally defined as the reciprocal of the free-space wavelength of light. Thus, once an optical frequency is traceable to the atomic time standard, it is possible to make a metrological transfer of the length standard to physical artefacts. In practical terms, frequency is transduced to length by interferometry.

Interferometry has been widely used in experimental checks of the most fundamental properties of the Universe due to its unmatched precision with the Michelson–Morley experiment being an example of its absolute success. Since the pioneering works of Einstein, research of gravitational waves has been purely theoretical for a very long time. However, laser interferometry turned out to have enough precision to detect even the smallest displacements necessary to detect gravitational waves. Chapter 11 "Gravitation measurements with laser interferometry" starts from the details of LIGO detector and proceeds to even more advanced approaches to experimental studies of gravity.

The introduction in the early 1960s of pulsed lasers for range-resolved measurements led to significant improvements in studies of the Earth and of the atmospheric environment. The optical and radar tracking of orbiting satellites at that time was incapable of yielding tracking-station coordinates with the accuracy of a few centimetres necessary for studies of processes such as Earth tides and plate tectonics. The satellite laser ranging (SLR) approach proposed by Plotkin [3], involving retro-reflectors on orbiting satellites, was subsequently shown to be capable of providing such accuracy. Chapter 12 is concerned with developments in SLR.

REFERENCES

1. Williams D C (ed) 1993 *Optical Methods in Engineering Metrology* (London: Chapman and Hall).
2. Culshaw B and Dakin J (ed) 1996 *Optical Fiber Sensors* vols 1–4 (Boston, MA: Artech House).
3. Plotkin H 1964 S66 laser satellite tracking experiment *Conf. Quantum Electronics* vol III (New York: Columbia University Press) pp. 1319–32.

2

Fundamental Length Metrology

Jens Flügge, Stefanie Kroker, and Harald Schnatz

CONTENTS

2.1 Introduction ... 3
2.2 Basics ... 3
  2.2.1 Evolution of the Metre – Definition and Realization ... 3
  2.2.2 Laser Interferometry ... 4
  2.2.3 Homodyne Laser Interferometry ... 4
  2.2.4 Heterodyne Laser Interferometry ... 5
  2.2.5 Interferometer Set-ups ... 7
  2.2.6 Grating Interferometers ... 7
2.3 Frequency Stabilized Lasers ... 8
  2.3.1 He-Ne laser Stabilized to the Gain Profile ... 9
  2.3.2 Iodine-Stabilized He-Ne Laser at $\lambda = 633\,\text{nm}$ ... 10
  2.3.3 Stabilized Frequency-Doubled Nd:YAG Laser at 532 nm Wavelength ... 11
  2.3.4 Frequency-Stabilized Diode Lasers ... 13
2.4 Practical Issues ... 14
  2.4.1 Refractometry ... 14
  2.4.2 Interpolation ... 14
  2.4.3 Accuracy Limits of Laser Interferometers ... 15
  2.4.4 Applications of Laser Interferometers ... 16
2.5 Multiple Wavelength Interferometry ... 18
  2.5.1 Gauge Block Calibration ... 18
  2.5.2 Interferometric Distance Measurements ... 19
References ... 19

2.1 Introduction

This section describes the use of lasers for length measurement traceable to the definition of the metre, which is defined today via the speed of light. High-precision and traceable length measurements are especially important for production engineering with a measurement range of up to some ten metres. Today for traceable measurements the use of laser interferometers relying on frequency-stabilized lasers is state-of-the-art. Laser interferometers and grating scales with interferometric readout are also widely used in industrial applications.

The chapter is divided in four main parts. 2.1 describes at first the definition and the history of the unit of length. Then the working principle of different laser interferometer types is sketched. Different methods for the frequency stabilization of lasers are explained in 2.2. The third Chapter 2.3 describes the possible error sources of laser interferometric length measurements and discusses advances and problems of laser interferometers in industrial use. In the last Chapter 2.4 multiple wavelength interferometers for the measurement of the length of gauge blocks and distances are briefly presented.

2.2 Basics

2.2.1 Evolution of the Metre – Definition and Realization

At the end of the last century an international treaty, the metre convention, sets the basis for the international system of units. At that time the metre was defined and realized as the distance of two lines on an international metre prototype made as a x-shaped bar of platinum-iridium at the temperature of melting ice, i.e. 0°C. This definition lasted from 1887 until 1960. The disadvantage of this definition resulted from the dependence of the unit on one single prototype, which was stored at the Bureau International des Poids et Mesures (BIPM). Another disadvantage was the limited measurement uncertainty in determining the position of single structures with photoelectric microscopes. Thanks to the progress of interferential length measurement and progress in the spectroscopy of gas discharge lamps, the metre was redefined as a multiple of the wavelength of a defined transition in ^{86}Kr in 1960. At that time, the invention of the gas laser gave an enormous scientific and technological input into further progress in the accuracy and

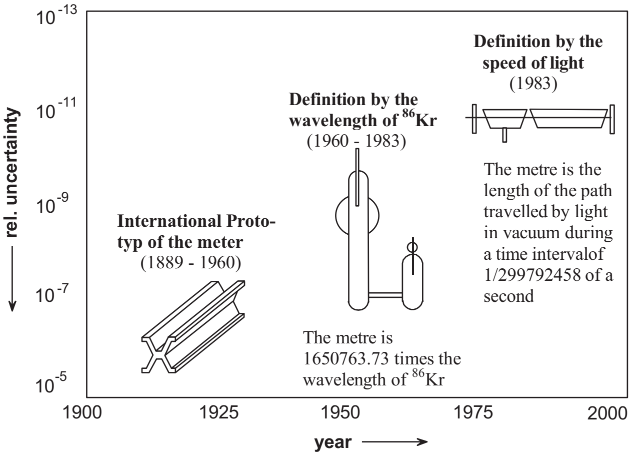


FIGURE 2.1 Definition of the SI-unit of length and their history.

the availability of light sources for length metrology. In 1983, the current definition of the metre as unit for length was determined by means of the speed of light [1] (see Figure 2.1).

This actual definition intimately links the unit of length to that of time and frequency. The velocity of light in vacuum is now a defined constant without an assigned uncertainty. However direct measurement of the time-of-flight of a light pulse is not possible for short distances with an adequate high precision. Time-of-flight measurements are mainly used in geodesy and astronomy, where large distances are measured with uncertainties in the range from millimetres to centimetres. Therefore, interferometry has to be used to transfer the unit of length to most industrially used dimensional artefacts, like line scales or gauge blocks, and measurement systems like linear encoders. Shortly after their invention in 1961, frequency-stabilized He-Ne lasers with their higher coherence length and higher optical power have replaced the spectral lamps for interferometric length measurements. They are today supplemented by other types of lasers. The vacuum wavelength λ_0 of the laser can directly be calculated from the frequency f and the velocity of light in vacuum c using the fundamental equation $\lambda_0 \times f = c$.

The unit of time is realized as 9192631770 times the period of the radiation due to the transition between two defined ground states of ^{133}CS with a relative uncertainty of a few 1e-16 [2,3]. Historically, the frequency of stabilized lasers in the visible spectral range has been determined by so-called frequency chains [4,5].

A novel development to replace the complex frequency chain is based on a femtosecond Titanium:Sapphire laser. It can be used for phase-coherent measurement of the frequency of an optical frequency standard with atomic clock accuracy. In the frequency domain, the pulse sequence of a femtosecond laser corresponds to a comb of frequencies which are separated by the pulse-repetition frequency. The width of this comb, which is given by the width of a single pulse, is spectrally broadened in a novel type of optical microstructure fibre to cover the entire visible and near-infra-red range. The frequency of any one of the comb lines can be obtained by measuring the line separation, i.e. the pulse-repetition rate, and a second frequency that gives the position of the entire comb with respect to the frequency origin. This gives a dense grid of millions of accurately known reference frequencies which can be used as a universal, self-referenced frequency ruler throughout the entire visible and near-infrared spectrum [6–8].

For practical length measurements it is preferable to use realizations of the metre based on lasers stabilized to atomic or molecular references. A list of frequencies and associated uncertainties of such stabilized lasers under defined conditions can be found in the "Mise en Pratique of the Definition of the Metre" [1] and its various updated versions that have since been approved by the CIPM [https://www.bipm.org/utils/en/pdf/mep_bibliography.pdf]. For example, for the most frequently used iodine stabilized He-Ne laser at 633 nm when operated under specified conditions, given in the Mise en Pratique, the relative frequency uncertainty of 2.5×10^{-11} is stated. However, the specific measurement conditions are by themselves insufficient to ensure that the stated standard uncertainty will be achieved. It is also necessary for the optical and electronic control systems to be operating with the appropriate technical performance. Nevertheless, the use of such recommended frequency standards for practical length measurement is generally sufficient.

2.2.2 Laser Interferometry

Today usually interferometric principles are used for practical length measurements of high precision. The first interferometric length measurements have been carried out by Michelson about one hundred years ago. But at that time the small coherence length and the low intensity of gas lamps made interferometric measurements a length measurement technology of only a few laboratories. The invention of the laser and the stabilization techniques especially for the He-Ne laser in combination with the development of opto- and microelectronics [9] resulted in an increasing number of applications.

Interferometric length measurement can be divided into displacement measurement and the measurement of absolute distances, e.g. for determination of the length of gauge blocks. The latter are described in some detail in Section 2.5.1. Displacement measurement is of high importance in industry. It is used in coordinate measuring machines, machine tools, and many other high-precision length measuring instruments. Today, the highest demands regarding the measurement accuracy arise from the manufacturing of integrated circuits, where interferometers and interferometric encoders are used to control wafer scanners for the exposure of the integrated circuits and mask metrology tools for reticle inspection. If highest accuracy is a matter of concern, interferometric techniques are the adequate choice.

Most interferometers for length measurements are based on the Michelson type. The detection principle of commercial laser interferometer systems can be distinguished between homodyne and heterodyne techniques.

2.2.3 Homodyne Laser Interferometry

The intensity I of an interference signal measured by a photodetector [10] is proportional to

$$I(\Delta s) \sim \tfrac{1}{2}E_M^2 + \tfrac{1}{2}E_R^2 + 2E_M E_R \cos(2\pi\Delta s_{\text{opt}}/\lambda) \qquad (2.1)$$

where E_M and E_R are the amplitudes of the electrical field of the measuring and the reference beam and $\Delta s_{\text{opt}} 2 \cdot n \cdot \Delta s$ is the optical path difference between these two interfering beams. Apparently the signal is unique only in a period of $\lambda/4$ for the position of the measurement reflector. To measure larger displacements, it is necessary to count the number of periods due to the zero crossings of the cosine term. However, with this signal alone, it is not possible to determine the direction of the reflector displacement. The first applied technology for detecting the direction made use of the homodyne principle by generating a second interference signal with a constant 90° phase shift [11].

In Figure 2.2 one possible optical set-up of a homodyne interferometer [12] is shown. The linear polarized laser beam is split into a measurement beam and a reference beam. In the measurement beam a $\lambda/4$ wave plate with the principle axes rotated 45° relative to the polarization of the beam produces a circular polarization state, where the perpendicular polarization states have a phase shift of 90°. In contrast, in the other beam these two perpendicular polarization states are in-phase. After superposition of the measurement beam and the reference beam a polarizing beam splitter (PBS) is used to generate two 90° phase-shifted interference signals. The direction of movement can now be determined at the zero-crossing of the interference signal using the sign of the other signal (compare right figure of Figure 2.2).

Additional signals phase-shifted by 180° and 270° phases are generated to remove the offsets $\frac{1}{2}E_M^2 + \frac{1}{2}E_R^2$ from the interference signals. The optical intensities are converted to a voltage with a trans-impedance amplifier and the voltages of each pair of 180° two signals can then be subtracted. The counting of the zero-crossings of both interference signals only gives a resolution of $\lambda/8$, which is often not sufficient for high-precision measurements. In those cases, it is necessary to enhance the resolution by interpolation [13]. The phase ϕ of the interference signal can be determined from the voltages of the 0° and the 90° signal by $\phi = \arctan(I_0/I_{90})$. The most common way is to convert the intensities with an AD-Converter and to use a digital logic based on the CORDIC algorithm for the fast calculation of the arctan function.

Another possibility to generate phase-shifted signals in a homodyne interferometer is to slightly tilt the reference mirror, which generates a spatial phase distribution in the resulting interferogram. If the photodetectors are positioned at different points of the interferogram, a phase difference of the measurement signals is generated. Often three 120° phase-shifted signals are used which can be analogue or digitally converted into two 90° phase-shifted signals, while also correcting for offset changes in case of laser intensity fluctuations [13A].

2.2.4 Heterodyne Laser Interferometry

The heterodyne principle [14] was first used in communication systems for high-sensitive detection. In a heterodyne interferometer, the interfering measuring and reference beams must have slightly different frequencies. The interference signal then consists of the sum frequency $f_1 + f_2$ and the difference frequency $\Delta f = f_1 - f_2$, also called the beat frequency. Photodetectors can only detect frequencies of up to a few gigahertz and therefore only the beat frequency. An example of a heterodyne interferometer [15,16] is shown in Figure 2.3. The two frequencies are separated by their polarization state, so such that a PBS can generate a measurement beam with the frequency f_1 and a reference beam with f_2. A movement of the measurement reflector with the velocity v causes a frequency shift $\delta f = 2 v f_1 / c$ in the measurement beam due to the Doppler effect. With a polarizer at 45°, the two signals interfere and a difference frequency of $\Delta f_M = f_1 \pm \delta f - f_2$ is measured. This signal is to be compared with a fixed reference frequency $\Delta f_R = f_1 - f_2$ measured in front of the polarizing beam splitter. The displacement Δs is then given by

$$\Delta s = \int_{t_1}^{t_2} v\,dt = \frac{\lambda_1}{2} \int_{t_1}^{t_2} \delta f_1\,dt = \frac{\lambda_1}{2} \int_{t_1}^{t_2} (\Delta f_M - \Delta f_R)\,dt. \tag{2.2}$$

When counting the periods of Δf_M and Δf_R simultaneously, their difference is proportional to the displacement. One counter digit represents a movement of $\lambda_1/2$. The direction of the movement can be determined directly from the sign of the difference. The measurement principle requires that δf must be smaller than Δf_R.

The heterodyne principle is very popular in the He-Ne-laser interferometry because the good signal-to-noise ratio enables to supply more than ten interferometer axes with one laser, even at high measurement speeds. Both the Zeeman- [17] and the two-mode stabilization technique [18,19] generate automatically two perpendicular polarized beams with different

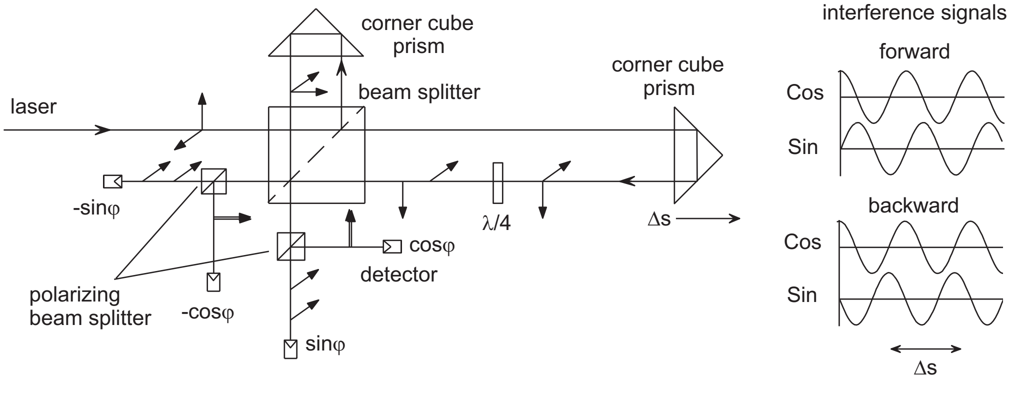


FIGURE 2.2 Principle of a homodyne laser interferometer.

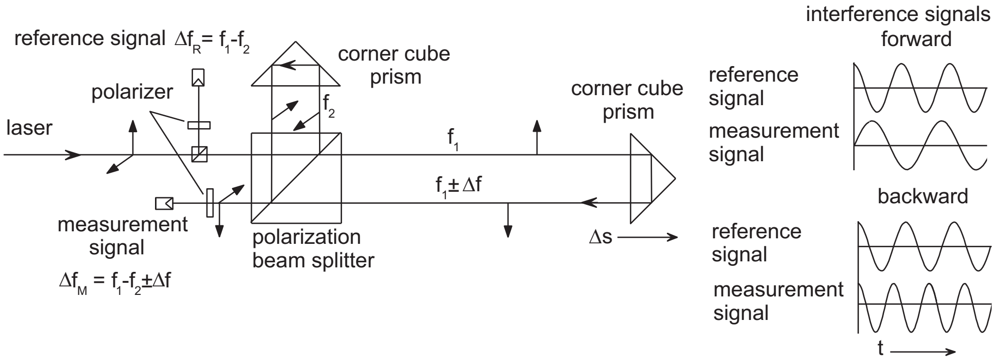

reference signal $\Delta f_R = f_1 - f_2$
polarizer
laser
measurement signal
$\Delta f_M = f_1 - f_2 \pm \Delta f$
corner cube prism
f_2
f_1
$f_1 \pm \Delta f$
polarization beam splitter
corner cube prism
Δs
interference signals
forward
reference signal
measurement signal
backward
reference signal
measurement signal
t

FIGURE 2.3 Principle of a heterodyne laser interferometer.

frequencies (see also 2.2.1). A Zeeman-stabilized laser has a beat frequency of a few megahertz, which limits the maximum displacement speed, while interpolation accuracy is harder to achieve at the beat frequency of two mode-stabilized lasers in the range of 500 MHz. Another possibility to generate a frequency difference is the use of external acousto-optic frequency modulators (AOM) [16].

The interpolation in a heterodyne system is equivalent to the measurement of the actual phase difference between Δf_M and Δf_R. With a fixed reference frequency Δf_R, this is equivalent to a measurement of the time between the zero-crossings of Δf_M and Δf_R. Examples of practical realizations implementing time measurements are given in Ref. [16]. Using the zero-crossings of the signal for phase determination based on time measurement is easy to be implemented. However, it has the disadvantage that the exact time of the zero-crossing will strongly be influenced by noise. Alternatively, the measurement signal can be demodulated by mixing with the reference signal. Mixing with a 90° phase-shifted reference generates a likewise 90° phase-shifted interference signal. This pair can be further handled in the same way as a homodyne signal pair. While this technique has already been used with analogue mixers, with the progress of digital signal processing, stable signal interpolation down to a few picometres [20–26] can be realized.

Another form of heterodyne interferometers employs a phase modulation in one arm of the interferometer. This can be achieved, for example, by electro-optic phase modulators (EOMs) [27]. With the progress of Distributed Feedback (DFB) diode lasers diodes especially for infrared wavelength typically used in telecommunication applications, some companies developed interferometers using wavelength modulation in unbalanced interferometers to generate the phase modulation [28]. This interferometer type can be set-up with optical fibre couplers to achieve very small sensor heads [29]. An example of such interferometer design with wavelength modulation can be seen in Figure 2.4.

The modulation depth Φ i defines the amplitude of the modulated phase as defined in formula 2.3. The modulation depth can be adjusted by the tuning range of the diode laser and the arm length difference in the interferometer Δs_{opt}. Analogue to formula 2.1, the intensity of a phase-modulated interferometer with the substitutions $C = \frac{1}{2}E_M^2 + \frac{1}{2}E_R^2$, $A = 2E_M E_R$, and $\phi = 2\pi \Delta s_{\text{opt}}/\lambda$ can be written as:

$$I(t) \sim C + A\cos(\phi + \varphi\cos(\omega_M t + \psi)) \tag{2.3}$$

After expanding, this function can be expressed as:

$$I(t) \sim C + AJ_0(\varphi)\cos\phi + \sum_{n=1}^{\infty} 2AJ_n(\varphi)\cos\left(\phi + n\frac{\pi}{2}\right)\cos(n(\omega t + \psi)) \tag{2.4}$$

where J_n is the Bessel function of the first kind of n-th order, which is graphically shown in Figure 2.4. When the interference signal has been amplified and converted to voltage, the harmonic signals can be separately demodulated using digital bandpass filters. The resulting voltages for the modulation frequency and for the first harmonics are:

$$\begin{aligned} V_1 &= -2AJ_1(\varphi)\sin(\phi) \\ V_2 &= -2AJ_2(\varphi)\cos(\phi) \end{aligned} \tag{2.5}$$

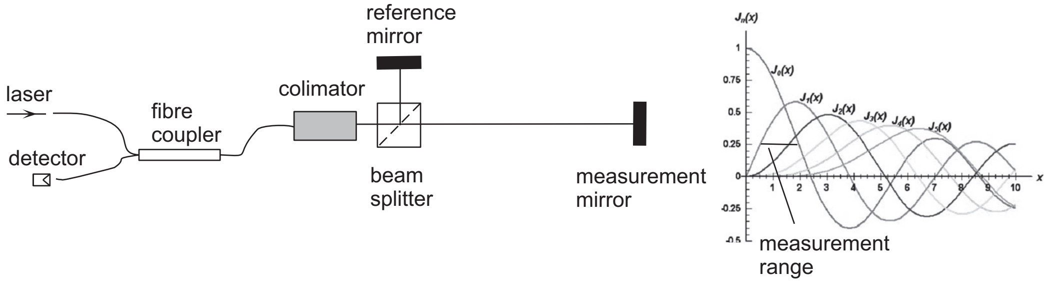

laser
detector
fibre coupler
colimator
reference mirror
beam splitter
measurement mirror
$J_n(x)$
1
0.75
0.5
0.25
0
-0.25
-0.5
$J_0(x)$
$J_1(x)$
$J_2(x)$
$J_3(x)$
$J_4(x)$
$J_5(x)$
1
2
3
4
5
6
7
8
9
10
x
measurement range

FIGURE 2.4 Principle of a wavelength-modulated heterodyne laser interferometer (left) and the Bessel functions of the first kind indicating the measurement range, where the signal intensities are larger than 25% (right).

These are similar signals to the homodyne interferometer again, but it is necessary to compensate the position-dependent ratio of the Bessel functions J_1 and J_2 before calculating the phase by the arctan function. The measurement range of the interferometer is limited to the sector, where J_1 and J_2 are over a threshold which is dependent on the noise level of the signals. This interferometer design is an example, where more complex optics designs can be exchanged against signal processing performance, whose price constantly decreases.

2.2.5 Interferometer Set-ups

In the classical laser interferometer design, corner cube prisms are used instead of plane mirrors [25,26], because this configuration is less sensitive to tilt during movement and it minimizes back-light to the laser, which can change the frequency of the laser source (see Figure 2.2). For two- or three-dimensional measurements and special set-ups, when plane mirrors are necessary, so-called double-plane mirror interferometers [27–29] are widely used. A sophisticated optical scheme is shown in Figure 2.5 as an example. Due to the corner cube prism, measurement beam and reference beam are always parallel to the incoming beam; therefore, the power of the interference signal remains more stable during tilt of the plane mirror. This "differential" interferometer measures the path difference between the measurement and the reference mirror with the same effective optical axis. Therefore, the position of all other optic components is of minor importance, simplifying the mounting of the interferometer components.

2.2.6 Grating Interferometers

For machine tools and coordinate measurement machines, normally encoders are used for length measurements, because interferometers are sensitive to refractive index variations, mainly caused by temperature and pressure fluctuations, and more expensive than encoders. With interferometric reading principles [30] introduced some years ago, resolutions comparable with laser interferometers can be achieved with such interferometric encoders, also known as grating interferometers. The principle of a grating interferometer is shown in Figure 2.6.

The intensity and phase of a beam diffracted by a grating are given by the Fourier coefficient of the diffraction order p. Due to the shifting theorem of the Fourier transformation $F\{\mathrm{f}(x-\Delta\mathrm{s})\} = e^{-i\cdot 2\pi p\Delta s/g} F\{\mathrm{f}(x)\}$, moving of the grating only affects the phase of the diffracted light, not its intensity [31]. Thus the phase shift Φ of the diffraction order p is $\Phi = 2\pi \cdot \mathrm{p} \cdot \Delta s/g$, where g is the period of the grating. The interference of different diffraction orders p and q generates the same interference signal as the homodyne interferometer described. earlier. The intensity as a function of the displacement is given in Equation 2.6.

$$I(\Delta s) \sim \tfrac{1}{2}E_p^2 + \tfrac{1}{2}E_q^2 + E_p E_q \cos\left(2\pi(p-q)\,\Delta s/g\right) \qquad (2.6)$$

The phase shift of the two interfering beams can result from the movement of grating as described above and from changes between the optical paths of the beams. Because the optical paths of the beams have nearly the same length, a change of the wavelength does practically not influence the phase between the beams. Therefore, no stabilization of the laser wavelength is necessary and standard laser diodes (LDs) can be used, which are cheap and emit enough optical power for achieving a high signal-to-noise ratio.

A special form of a grating interferometer is the so-called X-ray interferometer [32]. Three thin lamellas of silicon with (220) orientation are used in transmission as measurement grating and as well as beam splitter and combiner. An X-ray beam adjusted to the Bragg angle generates the X-ray interference. Due to the small grating period of about 192 pm, a resolution in the range of a few pm can be achieved.

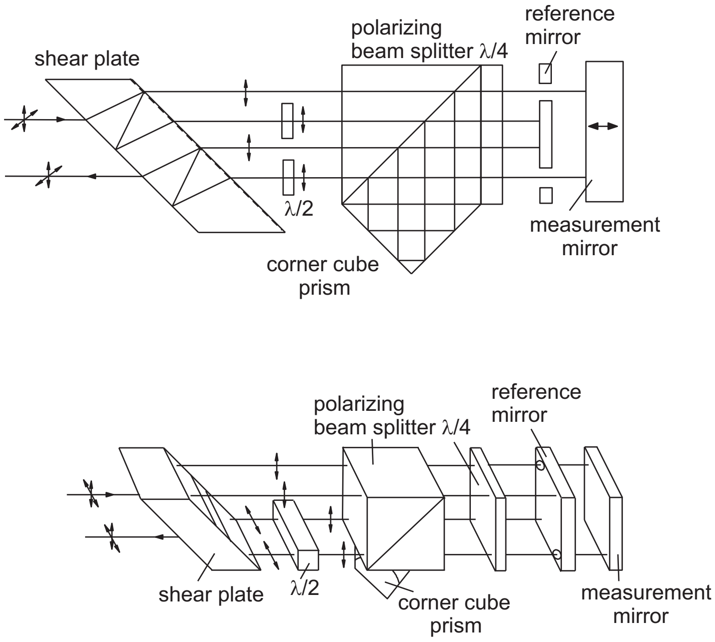


FIGURE 2.5 Principle of a double-plane mirror interferometer [28] and below a 3D sketch of the spatial layout.

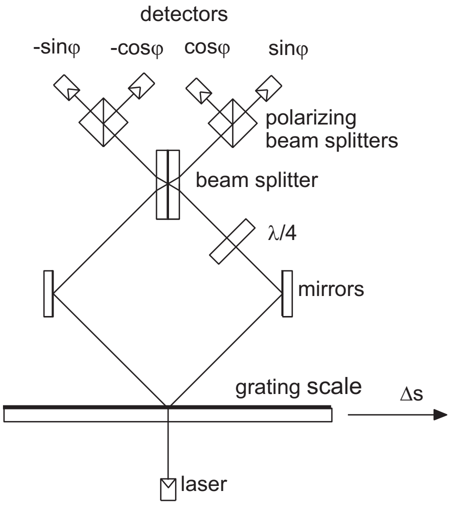


FIGURE 2.6 Principle of a grating interferometer.

2.3 Frequency Stabilized Lasers

The operating frequency or wavelength of any laser is determined by the optical path length of the laser resonator. Hence the wavelength of a laser in general is not constant but is directly coupled to the influences of the environment as, for example, temperature, pressure, or vibrations. The frequencies of microscopic references, however, like atomic, ionic, or molecular absorbers are almost free of these environmental influences. Hence, a laser can be used as a standard for dimensional metrology only if its wavelength or frequency is stabilized, for example, to suitable reference frequencies of atoms, molecules, or ions. If the laser frequency is tuned across an atomic resonance, the absorbed or emitted power varies. A suitable stabilization circuit converts this absorption or emission signal to an error signal which is used to servo-control the laser frequency to the centre of the line. The centre and width of the transition line are usually broadened and shifted, respectively, by different perturbations. For gaseous absorbers, for example, with typical thermal velocities of several hundred metres per second, the hierarchy of the broadening results from the first-order Doppler effect (~GHz), collisions with other absorbers or background gas (~MHz), the limited interaction time (~0.1 MHz), the recoil splitting (~1–100 kHz), or the second-order Doppler effect resulting from the time dilatation (~kHz). Furthermore, the frequencies differ for different isotopes (~GHz) and magnetic or electric fields may split and shift the absorption lines.

Depending on the intended application, the construction of frequency or wavelength standards follows two different philosophies. If the simplicity of use and operation is the primary goal, then one starts with an easy-to-handle laser whose frequency can be tuned only in a very limited range. Examples include the well-known He-Ne gas laser or Nd:YAG lasers. In this case, only absorption lines coinciding with the narrow emission ranges of these lasers can be selected as reference lines for the stabilization. Such absorption lines are available, for example, in the dense molecular spectrum of iodine in the visible spectral range (see, e.g. the iodine atlas of Gerstenkorn and Luc [33]) and in absorption lines of C_2H_2, CH_4, CO_2, SF_6, and OsO_4 in the infrared spectral range. The second philosophy avoids this restriction to accidental coincidences between laser lines and absorption lines by choosing an "ideal" absorption line. The higher accuracy achievable with this approach, in general, has to be paid for by the use of widely tuneable, coherent light sources such as dye lasers, diode lasers, optical parametric oscillators, and the implementation of extra methods to operate these sources in single mode and stable conditions despite the wide tuning range. Today, virtually any suitable absorption line in the optical range from the UV to the IR can be addressed allowing the selection of an ideal frequency reference for the corresponding application.

Frequency-stabilized lasers of both categories can be found in a list of recommended radiations for the realization of laser wavelength or frequency standards for the realization of the metre (see Table 2.1). This list is maintained and updated by the

TABLE 2.1

Optical Reference Wavelengths/Frequencies Recommended by the CIPM for the Realization of the Metre and for Scientific Applications.

Atom/molecule	Transition	Wavelength (fm)	Frequency (kHz)	Rel. Std. Uncer-tainty	Reference
H-atom	1S–2S (two photon)	243 134 624.626 04	1 233 030 706 593. 550	2×10^{-13}	[1]
^{171}Yb$^+$ ion	$^2S_{1/2}$–$^2F_{7/2}$	466 878 132.108 614	642 121 496 772. 645 0	6×10^{-16}	[137]
^{171}Yb atom	$6s^2\ ^1S_0$–$6s6p\ ^3P_0$	578 419 628.009 963	518 295 836 590. 863 6	5×10^{-16}	[137]
I_2-molecule	43–0, P(13), a_3	514 673 466.4	582 490 603 370	2.5×10^{-10}	[138]
I_2-molecule	32–0, R(56), a_{10}	532 245 036.104	563 260 223 513	8.9×10^{-12}	[1]
I_2-molecule	26–0, R(12), a_9	543 516 333.1	551 579 482 960	2.5×10^{-10}	[138]
I_2-molecule	9–2, R(47), a_7	611 970 770.0	489 880 354 900	3×10^{-10}	[138]
I_2-molecule	11–5, R(127), a_{16} (f)	632 991 212.58	473 612 353 604	2.1×10^{-11}	[1]
Unstabilized		632 990 800		1.5×10^{-6}	[139]
I_2-molecule	8–5, P(10), a_9	640 283 468.7	468 218 332 400	4.5×10^{-10}	[138]
^{40}Ca-atom	$^3S_0$–$^3P_1$	657 459 439.291 67	455 986 240 494. 150	1.1×10^{-13}	[1]
^{88}Sr$^+$-ion	$5^2S_{1/2}$–$4^2D_{5/2}$	674 025 590.8631	444 779 044 095. 486 5	1.5×10^{-15}	[137]
^{87}Sr atom	$5s^2$ 1S0–5s5p $^3P_0$	698 445 772. 516 39	429 228 004 229. 873 0	4×10^{-16}	[137]
^{85}Rb-atom (2-photon)	$5S_{1/2}$ (F = 3)–$5D_{5/2}$ (F = 5)	778 105 421.23	385 285 142 375	1.3×10^{-11}	[1]

International Committee of Weights and Measures (CIPM). The iodine-, methane-, or osmium tetraoxide-stabilized gas lasers belong to the first category, whereas the tuneable lasers stabilized to atomic hydrogen, calcium, rubidium, or to strontium and ytterbium ions belong to the second one. In general, the accuracy that can be achieved with the first category is sufficient for wavelength standards used in interferometric and dimensional measurements and hence are more widely used for reasons of simplicity and price. We therefore will mainly concentrate on this category in the following but will outline some of the newer development that may lead to future applications of the second one. Since the alignment of laser beams, the control of the interferometric quality of phase fronts as well as of diffraction is much easier with visible laser radiation, we will restrict ourselves to laser sources operating in this spectral range.

Besides the relative uncertainty of a stabilized laser corresponding to the accuracy of the laser and the capability to reproduce the metre in the SI, the stability of a stabilized laser is particularly important. The stability is a measure of the temporal frequency fluctuations or of the laser frequency noise. The frequency stability of lasers is often expressed in terms of the two-sample standard deviation being the square root of the Allan variance [34].

$$\sigma(2,\tau)/f = \frac{1}{f}\left\{\frac{1}{2\cdot(N-1)}\sum_{i=1}^{N-1}\left(\overline{\Delta f}_i - \overline{\Delta f}_{i+1}\right)^2\right\}^{\frac{1}{2}} \qquad (2.7)$$

The Allan standard deviation normalized to the frequency f is usually given for different sampling times τ. The two-sample relative standard deviation Δf_i can be calculated from the measured beat frequencies Δf_i integrated for different sampling times τ between two independently locked lasers.

2.3.1 He-Ne laser Stabilized to the Gain Profile

The gain curve of a He-Ne laser results from the Doppler broadening of the atomic laser transition and has a typical line width of about 1.5 GHz. The relative frequency uncertainty of a free-running He-Ne laser of $\Delta f/f = 1.5$ GHz/473.6 THz $\cong 3 \times 10^{-6}$ can be reduced if the laser frequency is kept fixed at a defined position of the gain profile. The two-mode-stabilized He-Ne laser uses the fact that the length of the resonator and therefore the axial mode separation can be adjusted such that over a large tuning range, only two adjacent axial modes oscillate. This is typically the case if the length of the resonator is about 30 cm corresponding to a mode separation of 500 MHz. Using laser tubes without apparent polarization-dependent losses, i.e. with internal mirrors rather than Brewster-angled mirrors, often these two modes are orthogonally polarized and are easily separated behind the rear mirror in a polarizing beam splitter [18].

When tuning the length of the resonator, both modes produce the signals at the respective photodetectors behind a polarizing beam splitter (Figure 2.7). If both signals are subtracted, e.g. in a differential amplifier, the resulting signal shows an asymmetric discriminant curve. If both photodetectors have the same sensitivity the resulting signal is symmetrical and shows a zero-crossing exactly at the atomic resonance. In the latter case the frequencies of both modes are symmetric with respect to the gain curve. If this zero-crossing where the intensities of both modes are equal is chosen as the reference point the difference signal represents an error signal. The antisymmetric shape of the error signal allows to discriminate whether the length of the frequency has to be increased or reduced to return to the reference point. Even though the two-mode stabilization technique is frequently used for He-Ne lasers operating in the red (λ= 633 nm) or green ($\lambda = 543$ nm) spectral range there are also stabilization techniques to the gain profile based on other techniques.

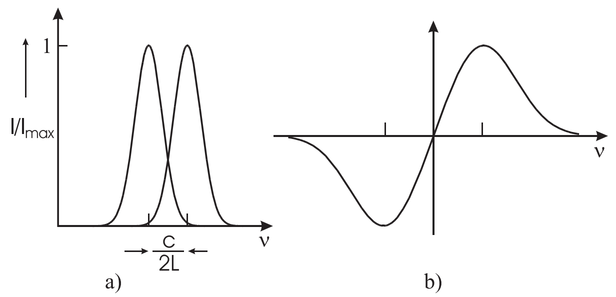


FIGURE 2.7 (a) Two adjacent but orthogonally polarized modes separated by a polarizer are detected by two independent photodetectors with their maxima shifted by the free spectral range *c/2L* of the laser resonator. (b) The difference signal of the photodetectors resulting in an antisymmetric error signal suitable for laser stabilization.

If the laser tube is placed into an axial magnetic field the energy levels of the neon atoms in the amplifying medium are shifted due to the Zeeman effect proportional to the applied magnetic field. Consequently, the laser line splits into two oppositely circularly polarized waves, the frequencies of which differ depending on the magnetic field by 300 kHz–2 MHz. The two circularly polarized waves are converted into two orthogonally linearly polarized waves by means of a quarter wave plate. Again, the difference between the powers of the two waves detected by the two detectors can be used to stabilize the laser frequency.

The strong absorption near the resonance of the neon atoms is accompanied by a strong dispersion which leads to a frequency-dependent change of the index of refraction in the vicinity of the line centre. As a result, the frequency of each one of the two Zeeman modes depends on their frequency difference from the line centre. It has been shown that the frequency difference between both Zeeman modes shows a minimum at the line centre which can be utilized to stabilize the laser frequency [35].

The different methods of the laser frequency stabilization to the gain profile meet with advantages and disadvantages. The simplicity of the two-mode stabilization with two orthogonally polarized resonator modes makes it useful for operation in interferometers, particularly in combination with heterodyne interferometers (see 2.1.3). Disadvantages result if it is necessary to suppress one mode and from the fact that the used mode is not in the centre of the gain profile. Depending on the polarity of the difference signal, the mode can be stabilized to either side of the gain profile. Furthermore, there are lasers as, for example, the green He-Ne lasers at 543 nm where the modes change their polarization when going through the centre of the gain profile. These polarization jumps make the two-mode stabilization technique not applicable and additional means may be necessary as, for example, by placing an additional magnet near the gain tube [36]. Moreover, the electronic lock point may be shifted by unequal photo diodes and electronic offsets.

When the Zeeman splitting is used for frequency stabilization, the frequency difference of both modes is much smaller than in the case of the two resonator modes. The advantage of the higher slope of the discriminant curve with the associated high gain available to the servo system is compensated by the reduced locking range.

Investigations of the frequencies of polarization-stabilized red He-Ne lasers during a period of more than two years showed a drift of about 5 MHz [37] corresponding to a fractional frequency variation of 10^{-8}, which is sufficient for nearly all practical length measurements. These lasers often show frequency variations of similar magnitude on external magnetic fields, temperature variations, and ageing due to pressure loss of the gain tube. Repeated calibrations of several lasers at PTB over longer periods showed similar results.

2.3.2 Iodine-Stabilized He-Ne Laser at $\lambda = 633$ nm

Molecular iodine has a rich spectrum of more than 50 000 absorption lines from the green to the red parts of the visible spectrum [33]. Hence, virtually any Doppler-broadened emission line of a gas laser covers several hyperfine lines that can be used for frequency stabilization. Following the early work of Hanes und Dahlstrom [38] several lasers have been stabilized. By far, the most widely used one is the He-Ne laser at $\lambda = 633$ nm, where the absorption frequency of the vibrational transition 11-5 of the R(127) line of the iodine isotope $^{127}I_2$ coincides with the emission frequency of the isotope ^{22}Ne in a He-Ne laser. These absorption lines, however, are not very strong, and in order to detect the absorption signals with high signal-to-noise ratio, a higher power would be necessary as the output power available in general from the He-Ne laser. Two orders of magnitude higher intensities can be obtained for an absorber contained in a cell inside the resonator (Figure 2.8).

Owing to the thermal velocity distribution of the iodine molecules in the absorption cell, the hyperfine lines of iodine are Doppler-broadened. For an arbitrary laser frequency f_L, both counter-propagating laser beams inside the resonator in general interact resonantly with different Doppler-shifted velocity groups. If, however, the frequency of the laser coincides with a transition frequency of the molecules at rest, both laser beams are interacting with the same velocity group of molecules having zero velocity in the direction of the laser beams. Correspondingly, the absorption of these molecules is reduced due to the saturation of the corresponding non-linear absorption. Hence, the absorption losses inside the laser resonator decrease and the output power of the laser increases.

For the conditions of the typical iodine-stabilized He-Ne laser at $\lambda = 633$ nm (see 2.9), the output power increases by only about 0.1%. To detect the absorption signal in the noise, the frequency of the laser is modulated, and the corresponding synchronous variation of the laser power is phase-sensitively measured. The laser frequency is modulated by a voltage with a frequency f of a few kilohertz which is applied to the piezoelectric transducer (lead zirconium titanate) attached to one of the laser mirrors. The phase sensitive detector (PSD) detects the signal from the photo diode which is in phase with the modulation signal and all other frequency components cancel. This method differentiates the signal and therefore subtracts a constant frequency-independent background. A suppression of a non-constant background can be achieved using higher order derivatives of the signal. The third-order derivative of the absorption signal, for example, is widely used since (as any other odd derivatives) it has a zero-crossing at the line centre and any linear and quadratic background is removed (see Figure 2.9). Stabilization schemes employing higher order derivatives have been utilized [39,40] and frequency shifts of up to 35 kHz have been observed when switching from the third to the fifth harmonic-locking.

The reproducibility of the iodine-stabilized He-Ne laser was thoroughly investigated and is well documented [41] mainly thanks to the large number of inter-comparisons between the several institutions operating these systems. The laser frequency depends on the various operational for example,

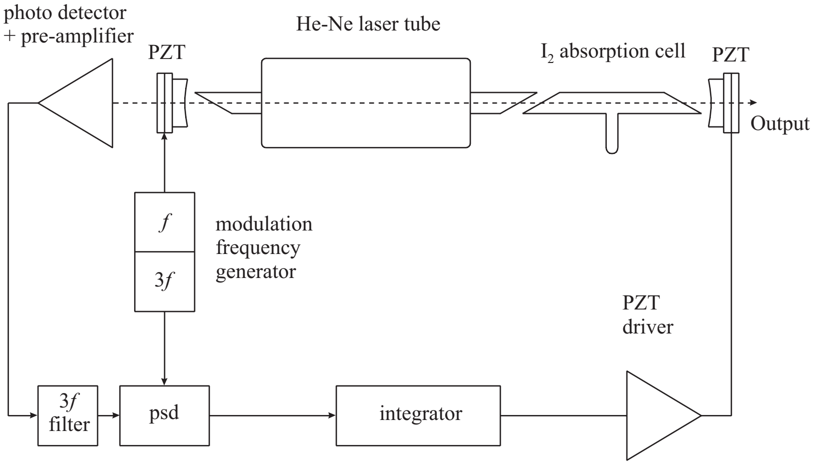


FIGURE 2.8 Schematics of an iodine-stabilized He-Ne laser with the absorption cell inside the laser resonator.

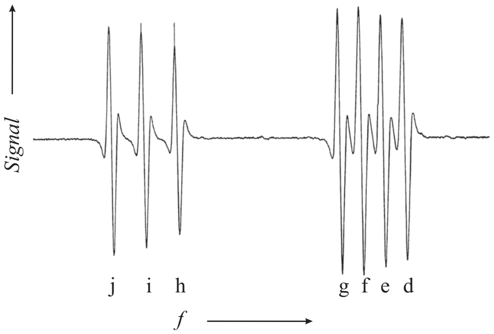


FIGURE 2.9 Third-harmonic spectrum of the iodine-stabilized He-Ne laser at the R(127)11-5 transition near 633 nm.

the modulation width, the vapour pressure in the absorption cell, or the intra-cavity laser power. Typical values for these dependencies are –10 kHz MHz^{-1}, 6 kHz Pa^{-1}, and –1 kHz mW^{-1}, respectively. International comparisons showed that the frequencies of the majority of the iodine-stabilized lasers at 633 nm agree to about 10 kHz if all lasers are operated under the same conditions that are recommended by the CIPM [1]. To meet these conditions, the cold finger and the wall of the iodine cell have to be kept at (15 ± 0.2)°C and (25 ± 5)°C, respectively. Furthermore, the full modulation width of the laser frequency and the intra-cavity laser power have to be restricted to (6.0 ± 0.3) MHz and to (10 ± 5) mW, respectively. For lasers operated under these conditions and good practice, the recommendation [1] gives a relative uncertainty of the frequency of the laser of 2.5 ± 10^{-11}.

The variation of the laser frequency on the modulation width results from the residual Doppler background in combination with a possible asymmetry of the absorption line. The temperature of the cold finger of the iodine cell is used to adjust the vapour pressure thereby affecting the rate and duration of the collisions with the associated pressure-broadening and shift of the absorption line. The variation of the laser frequency with the laser power depends on the particular design of the laser. In order not to exceed the relative uncertainty of 2.5×10^{-11}, this contribution must not be larger than 1.4 kHz mW^{-1}. This dependency results, for example, from the mixed influence of the saturation in the iodine cell and by a variation of the index of refraction in the discharge leading to a gas lens effect. These effects seem to limit ultimately the performance of the iodine-stabilized laser at 633 nm which despite these deficiencies can be expected to be the workhorse for some years to come. He-Ne lasers at other wavelengths, for example, at 543 nm have been stabilized with the third harmonic technique or the modulation technique to be described below [42]. In the following we discuss a candidate that at least to some extent may replace these He-Ne lasers.

2.3.3 Stabilized Frequency-Doubled Nd:YAG Laser at 532 nm Wavelength

Frequency-doubled Nd:YAG lasers (532 nm) pumped by diode lasers are of particular interest for applications as optical frequency standards in view of their high efficiency, high output power, and low intrinsic noise. Small lasers that provide us with output powers of 100 mW and more at 532 nm are commercially available. A part of their emission range coincides with strong absorption lines of molecular iodine [43] which have an order of magnitude smaller line width as compared to the ones used in the red He-Ne standard. The high power of the frequency-doubled Nd:YAG laser not only allows to serve several interferometers at the same time but also enables one to use external absorption cells. Due to the latter property there is no need for a modulation of the output power which in turn allows fast interferometric measurements.

Various methods have been used in different laboratories to generate the error signal. The methods of modulation transfer spectroscopy [44,45] and phase modulation spectroscopy [46,47] are very powerful to achieve discriminant signals of high signal-to-noise ratio. As an example, we discuss here a laser stabilized by means of the method of phase modulation spectroscopy since it allows for higher modulation frequency, and therefore, the influence of residual low-frequency technical noise of the commercial laser can be better suppressed.

Figure 2.10 shows the layout of the stabilization scheme as taken from Cordiale et al. [48]. The output beam of a commercial, frequency-doubled Nd:YAG laser (approximately 100 mW) passes through a telescope which forms a collimated laser beam of about 2 mm diameter. Typically, a few milliwatts of the output needed for the stabilization pass through a PBS 1, whereas the deflected beam is available for experimental use. The power ratio between the two partial beams can be varied by rotating the half-wave plate in front of PBS 1. PBS 2 then divides the laser beam into a pump (saturating) beam and a probe beam. Again, the power ratio between the pump and the probe is adjusted by the second half-wave plate. An AOM shifts the frequency of the pump beam by $f_{\text{AOM}} = 80$ MHz. The radio frequency (rf) signal driving the AOM is switched on and off at an audio frequency (AF) of 23 kHz corresponding to a fast chopping of the deflected (and frequency-shifted) saturating pump beam. The pump beam passes through the absorption cell and periodically saturates the absorption of those iodine molecules of which the Doppler-shifted transition frequency coincides with the frequency of the pump beam. The saturated absorption is then probed by a counter-propagating laser beam of the frequency f. The superposition of the counter-propagating beams (pump and probe) of different frequencies results in a "walking wave" structure where the nodes and anti-nodes move with a velocity $c \cdot f_{\text{AOM}}/2f$. Iodine molecules moving with the same velocity component experience a standing wave, and the rules of saturated absorption can be applied. Correspondingly, the laser frequency f at the centre of the observed saturation dip is Doppler-shifted by an amount of $f_{\text{AOM}}/2$ (to first order) and the transition frequency of the molecule at rest is given by $f_0 = f_{\text{L}} + f_{\text{AOM}}/2$. In order to probe the "saturation holes" generated by the pump beam, the phase of the probe beam is modulated by an EOM at a modulation frequency of 5 MHz. When the laser frequency is tuned through the molecular resonance, an intensity modulation at that frequency occurs which is detected by a photodetector (PD) and demodulated by a PSD to generate an error signal for the stabilization. Frequency offsets generated by a residual linear absorption are strongly suppressed by chopping the pump beam and phase-sensitive detection. Consequently, the

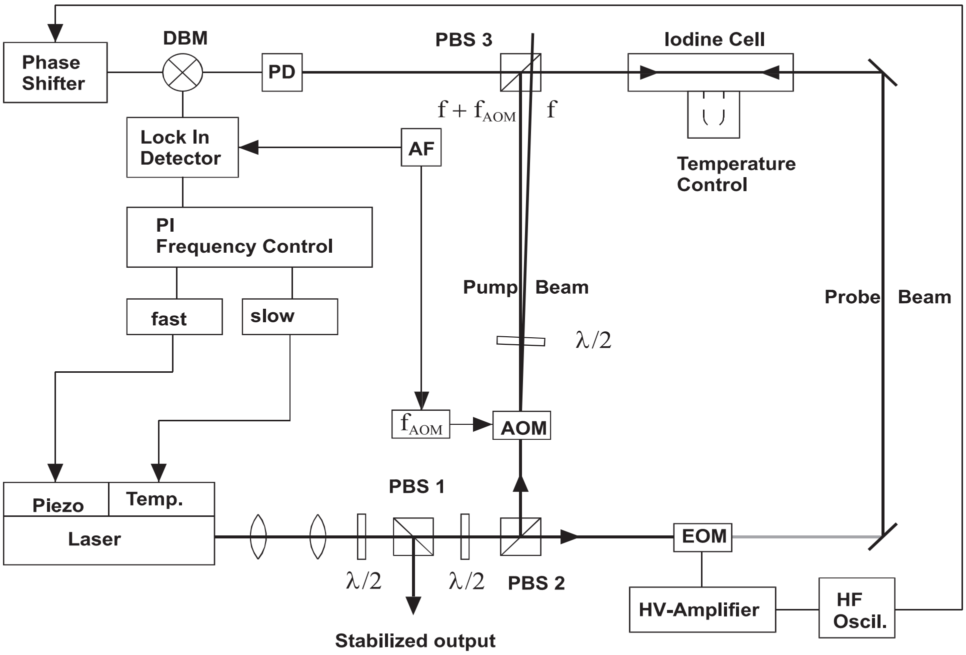


FIGURE 2.10 Layout of an iodine-stabilized Nd:YAG laser. DBM: double-balanced mixer; AF: audio-frequency generator; AOM: acousto-optic modulator; EOM: electro-optic phase modulator; PD: photodetector; PBS: polarizing beam splitter.

demodulated signal of the double-balanced mixer (DBM) is in turn phase-sensitively detected by a lock-in detector, which is driven with the chopping frequency (23 kHz). With this method, only the non-linear component, i.e. the saturated absorption, contributes to the error signal. Depending on the setting of the phase of the DBM, the in-phase or the quadrature component of the interaction with the iodine molecules is measured (see Figure 2.11).

The dispersive component has a steep slope with a zero-crossing at the iodine frequency that can be used to stabilize the laser. These lasers have been set up at various places (e.g. [45,48–50]). Comparisons of such lasers have been performed [48,50] with frequency differences between independent lasers of only a few kilohertz. Among other effects, the stabilized frequency critically depends on a precise alignment of the two counter-propagating waves in the iodine absorption cell, on a residual amplitude modulation generated by the EOM, and on spurious optical feedback within the optical set-up. In the case of modulation transfer spectroscopy, it has been demonstrated that a frequency reproducibility of a few hundred hertz can be achieved [45]. The instability as expressed by the relative Allan standard deviation of a few parts in 10^{13} [50] also makes this laser very attractive for rapid interferometric data acquisition. Regarding its high power, its small size, and its high frequency reproducibility, the iodine-stabilized, frequency-doubled Nd:YAG laser represents an optical frequency standard with important applications in precision length metrology, interferometry, and spectroscopy.

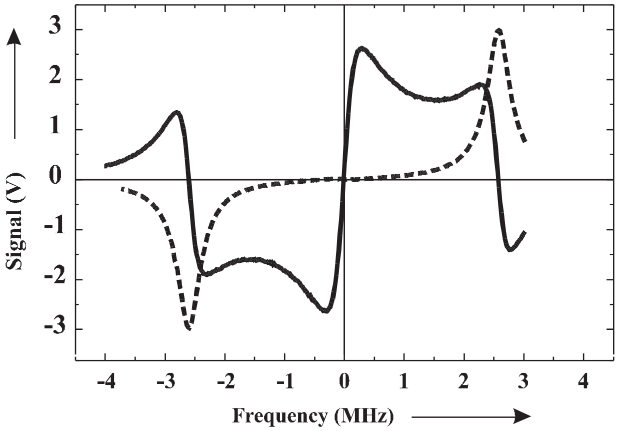


FIGURE 2.11 In contrast to the absorptive signal (- - -), the dispersive (----) signal from a hyperfine component of iodine can be used to frequency-stabilize a frequency-doubled Nd:YAG laser.

2.3.4 Frequency-Stabilized Diode Lasers

Diode lasers have the reputation to be cheap, compact, long-lived devices that can be easily operated over a large wavelength range of more than 10 nm. They are furthermore available for wavelengths from about 1.7 μm throughout the infrared to about 630 nm and recently also near 400 nm. Consequently, a great deal of work has been performed to develop diode-laser-based wavelength standards. As we will see, today's diode laser systems for length metrology are neither cheap nor simple since the benefit of the tuneability and small size has to be paid for by additional efforts to achieve the required stability and beam quality, respectively.

Tuneable diode lasers are predominantly useful if the power available from He-Ne lasers is not sufficiently high [51], or as discussed in the beginning, if the quality of the available absorption lines coinciding with the suitable fixed laser lines is not sufficient. As an example for the latter case, we mention the P(33) 6-3 transition or others in molecular iodine that have higher absorption as compared to the R(127)11-5 line that is used to stabilize the He-Ne laser but are not accessible to this laser. Other suitable absorbers include the strong Rb D2 lines near 780 nm [52,53] or the two-photon Rb transition at 778 nm [54] and the very narrow Ca transition at 657 nm of Table 2.1.

In the visible, however, commercially available diode lasers of the Fabry-Pérot type have line widths of several hundred MHz and can therefore not directly be used for dimensional interferometry due to the limited coherence length. This line width can be substantially reduced by employing an extended cavity diode laser (ECDL) configuration (Figure 2.13) [55,56]. In this configuration, one output facet of the diode laser is often anti-reflection-coated, and the feedback from this reflecting surface is replaced by an external reflector. Hence, the length and consequently the Q-factor of the cavity can be increased and allow one to even use frequency-selective elements. Typically, a grating is used as frequency-selective element (see Figure 2.12). The grating can be used in a Littrow or Littman [57,58] arrangement. In the former one the angle is chosen such that the incident beam is retro-reflected into the LD and the zero-order beam is coupled out. The direction of this beam varies when the grating is rotated in order to tune the wavelength. The Littman configuration avoids this disadvantage by rotating a feedback mirror (FB) and keeping the orientation of the grating fixed. The Littman configuration is shown in Figure 2.12.

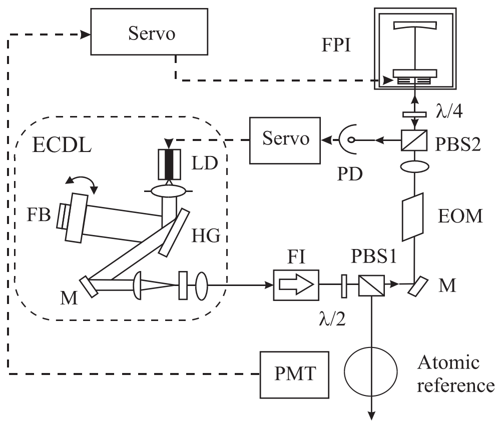


FIGURE 2.13 Set-up of an ECDL pre-stabilized to a FPI and long-term stabilized to an atomic reference. LD: laser diode; M: mirror; FB: feedback mirror; HG: holographic grating; FI: Faraday isolator; PBS: polarizing beam splitter; EOM: electro-optic phase modulator; PD: photo diode; $\lambda/4$: quarter wave plate; $\lambda/2$: half wave plate; FPI: Fabry-Pérot interferometer; PMT photomultiplier tube detecting the fluorescence light from the excited atomic reference.

Typical ECDLs use holographic gratings (HGs) with 1800 or 2400 lines/mm and a length of the resonator of a few centimetres. The free-running line width of such a laser is typically of the order of 100 kHz which is sufficient in most cases. For the laser frequency standards of the lowest uncertainty as, for example, the Ca-stabilized laser [59] (Table 2.1), the line width can be further reduced by employing a Fabry-Pérot resonator (FPI) and a phase modulation technique [54] similar to the phase modulation techniques used to stabilize the Nd:YAG laser described earlier. Part of the laser output is then used to stabilize the frequency by a suitable means to the atomic or molecular reference [60].

With the described set-up of a calcium-stabilized laser [59] a relative uncertainty of 1.3×10^{-12} was achieved which shows an order of magnitude improvement with respect to the iodine-stabilized He-Ne laser at 633 nm. At the same time the instability of $\sigma(2,\tau) < 10^{-12}$ at 1 s was also smaller by about an order of magnitude. Much effort is now devoted to the development of iodine-stabilized diode lasers at 633 nm [61–63]. A recent intercomparison of eight different diode laser standards showed for the first time a short-term frequency stability of 4×10^{-12} for 1 s, which is better than the classical He-Ne laser. Besides these

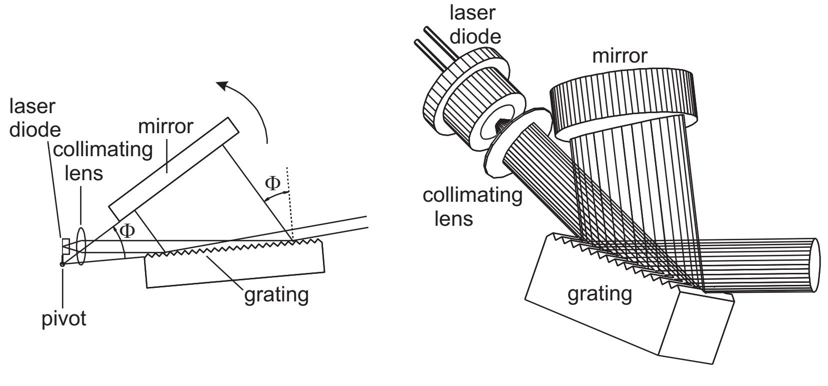


FIGURE 2.12 Principle of an external cavity diode laser in the Littman configuration and a 3D-scaled example.

high-end wavelength standards diode lasers may be stabilized by the opto-galvanic effect where the change in the electrical properties of a discharge lamp is detected when the discharge is illuminated by radiation with a frequency close to an atomic transition of the atoms in the discharge lamp [64].

2.4 Practical Issues

2.4.1 Refractometry

The actual wavelength λ effective in most laser interferometers is dependent on the refractive index n of air. The dependency is $\lambda = \lambda_0/n$. The knowledge of the refractive index is one of the main problems for precision interferometry under atmospheric conditions. The refractive index depends on wavelength, pressure, temperature, and the air composition. The following values give an approximation of the influence of the main environmental parameters on the refractive index. Values for other gas contents can be found in [67].

- temperature $\frac{\Delta n}{\Delta T} \approx -1 \cdot 10^{-6}\,K^{-1}$
- pressure $\frac{\Delta n}{\Delta p} \approx 2.7 \cdot 10^{-9}\,Pa^{-1}$
- humidity $\frac{\Delta n}{\Delta h_r} \approx -10^{-8}\,(\%)^{-1}$
- CO_2-content $\frac{\Delta n}{\Delta CO_2} \approx 1.5 \cdot 10^{-10}\,ppm^{-1}$

To minimize the effect of the refractive index, the interferometer can be operated in vacuum. Some high-precision tools like electron microscopes, electron-beam writers, or Extreme Ultra Violet (EUV) lithography tools anyhow work in vacuum. To avoid the effort of designing a complete vacuum tool, it is also possible to shield the interferometer measurement beam by a vacuum bellow [65,66], but a larger number of interferometers operate in air.

There are two main principles used to determine the refractive index. The first and mostly used principle is the indirect determination through the measurement of above parameters and calculation of n using Edlén's equation [68–70] which is based on interferometric measurements. The accuracy of Edlén's formula has been increased due to new experimental results of some metrological state institutes [71,72]. The pressure compensation occurs with the speed of sound; therefore, only one pressure gauge is necessary. Because the influence of humidity and CO_2 content are small enough, a single measurement system is also adequate. For longer distances, multiple temperature sensors should be used, because temperature gradients cannot be neglected in most practical cases. The advantage of this parameter measurement principle is that the set-up is easy, cheap, and reliable. The disadvantage is that changes in the gas composition cannot be detected and due to the time constant of the sensors fast fluctuations of the refractive index cannot be detected. The time constant of Pt 100 measurement resistors is about 30 s, while the response of the interferometer on changes of the refractive index is nearly instantaneous. Additionally, the temperature sensors can be affected by thermal radiation.

The second principle is the direct interferometric measurement of n using a refractometer. There are two types of refractometers, tracking refractometers which only measure changes of the refractive index and absolute measuring refractometers which use vacuum as a reference for the determination of the absolute refractive index. A tracking refractometer is in principle an interferometer with a fixed mechanical length of the measurement arm. Any change of the measurement value must be caused by the refractive index change, which can be calculated, if the mechanical length is known [73]. For higher precision changes of the mechanical length can also be measured by a second interferometer beam in vacuum tubes [16]. The absolute value of the refractive index must be initialized, if necessary, by an additional measurement system. Tracking refractometers are actually the only possibility to measure fast changes of the refractive index. But a tracking refractometer does not measure the refractive index in the measurement beam of the interferometer itself. Therefore, the performance of a tracking refractometer relies on the homogeneity of the environment.

The absolute value of the refractive index can be determined by a comparison with the wavelength in vacuum [74,75]. A possibility for the comparison is the interferometric measurement of the change of the optical length of a stable chamber during filling with air. For optimization of the refractometer, a second interferometer with the same measurement axis is necessary to compensate for length changes and bending of the chamber due to compressibility or temperature changes caused by ventilation. The principle of an absolute measuring refractometer is shown in Figure 2.14.

Due to the principle the measurement speed is low. The air temperature in a chamber normally differs from the air temperature in the interferometer beam, which requires an additional temperature correction. Such refractometers are expensive and there are nearly no practical advantages for industrial applications compared with the parameter method [76].

An interesting method for long measurement ranges is the use of two wavelengths in the interferometer to make it less sensitive against environmental conditions. With this method it is possible based on Edlén's formula to calculate the refractive index from the dispersion, the difference between the refractive indexes of the two wavelengths, which is less sensitive to pressure and temperature. The problem is that the resolution decreases proportional to $(n_1 - n_2)^{-1}$, where n_i is the refractive index at the wavelength λ_i. Thus, two wavelengths with a large difference in the refractive index must be used [77–79].

2.4.2 Interpolation

To enhance the resolution of an interferometer, the signal period of $\lambda/2$ can be interpolated as described in Chapter 2.1. The interpolation in most interferometers is not fully linear. Normally, a measurement deviation with the period of $\lambda/2$ and also with higher spatial frequencies occurs. The reason for the interpolation non-linearities is partly different for homodyne and heterodyne interferometers. In both interferometers, reflexions running multiple times through the interferometer and adding intensity at the detector are generating non-linearities in the interpolation [80]. This error cannot be corrected.

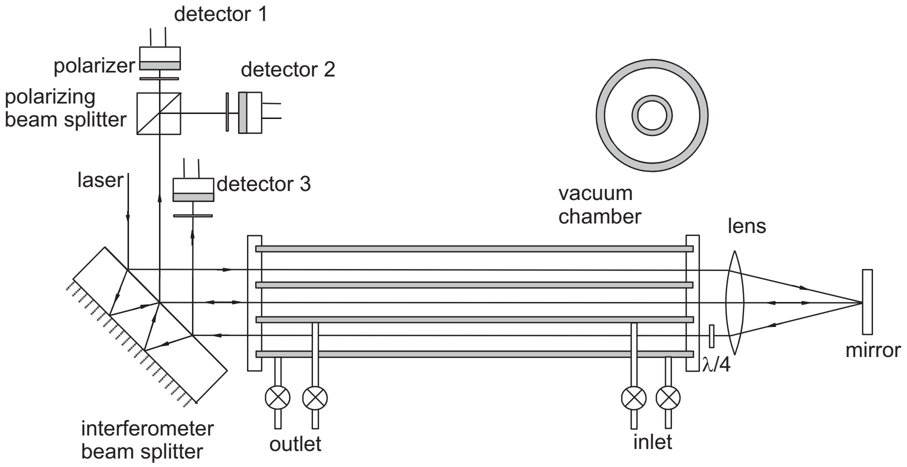


FIGURE 2.14 Principle of the absolute measuring refractometer set-up at NPL [75].

To prevent this type of reflexions the optical elements can be slightly tilted or interferometer optics without surfaces perpendicular to the beam can be used [66].

The interpolation in a homodyne interferometer assumes the two interference signals being ideal. That would mean the signals have exactly 90° phase shift, the same amplitude, and no offsets. These parameters depend on the quality of optics, alignment, and amplifiers. They can also change as a function of the measurement position. Typically 1% of the signal period is achievable. Different methods for the correction of non-linearity errors have been demonstrated. A widely adopted method was proposed by Heydemann [81,82]. He has presented a method to calculate from a series of previously measured intensity values over at least one signal period from a set of correction parameters by linear fitting. These parameters can then be used to correct the interference signals. The non-linearity of interpolation can be reduced down to $1\%_0$ of the signal period, or even less [83]. In heterodyne interferometers, the non-linearity is caused by mixing of the two frequencies in the interferometer beams [84,85]. The non-linearity has mainly the period of $\lambda/2$, but also components with a period of $\lambda/4$ can be observed [86,87]. The effect is caused, for example, by elliptical polarization of the laser beam [88] and by non-ideal beam splitters [89]. Every additional optical element, especially all types of optical fibres, in front of the beam splitter can increase the frequency mixing [90]. A detailed theory of the non-linearity in heterodyne interferometers can be found in [91]. The non-linearity has typically amplitude of about one to a few nanometres and is therefore comparable to the effect in homodyne interferometers. A possibility to reduce the non-linearity with a $\lambda/2$ plate and a second receiver is given in [92]. Another example of correcting non-linearities is shown in [93]. The effect of non-linear interpolation generated by frequency mixing can be fully removed if the two frequencies are spatially separated and not distinguished due to the polarization [94]. But in this case, the frequency offset must be generated outside of the laser, for example, with an AOM in one of the beams. Additionally the optical set-up is more complicated because a reference interferometer is necessary to compensate phase shifts which occur during the generation of the frequency offset, but it allows to set up a fully fibre-coupled interferometer nearly free from periodic non-linearity errors [26].

The interpolation and especially the accuracy of correction methods are limited by the noise of the interference signals [95]. The noise is mainly caused by the laser itself and the photodetector. Hence, laser sources of high output power like a frequency-doubled Nd-Yag laser or a fibre laser can reduce the noise in the interferometer. The noise in homodyne interferometers is theoretically larger than in heterodyne interferometers because the noise density is higher in the DC range. In practice due to the quality of today's DC amplifiers the performance difference is negligible but commercial heterodyne interferometers still allow operating a larger number of axes per laser. Actual limits for the interpolation are shown in Figure 2.15.

Even higher resolution can be achieved by Fabry-Pérot interferometers [96,97]. Resonators with high Q-factors generate a very sharp interference signal with much higher slopes compared to the signals of a Michelson interferometer. Therefore, noise on the signal causes smaller errors regarding length deviations. Unfortunately, the Fabry-Pérot interferometer can only be used, when the laser frequency corresponds to the resonance frequency of the resonator. This requires that the laser source must be constantly controlled to match the resonance. A length change can then be measured by determining the frequency difference of the laser between two positions. The frequency of the laser must be determined by a beat frequency measurement relative to another stabilized laser. Because this system is complicated and not well suited for larger measuring length or dynamic applications, it is in practice rarely used.

2.4.3 Accuracy Limits of Laser Interferometers

There are some other smaller effects often neglected. Measurement and reference beam should have the same optical path length in glass, to minimize thermal effects. In addition, variations in thermal gradients can cause small measurement errors. As the laser beam has no perfect plane wavefront and its intensity profile has a Gaussian shape, diffraction effects change the form of the phase front and beam diameter over the

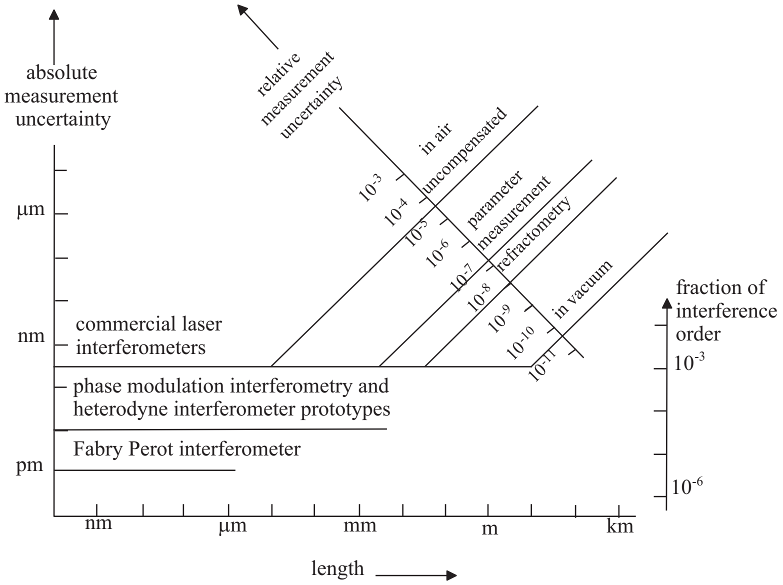


FIGURE 2.15 Accuracy limits for laser interferometry.

measurement length [98,99]. Therefore, the signal period of the real interferometer slightly differs from an interferometer with a hypothetic ideal plane wave. Because the wavefront can also be changed by optic components, for highest precision multiple path interferometers should be avoided. Figure 2.15 summarizes the discussed limits of accuracy for displacement laser interferometers [100,101].

2.4.4 Applications of Laser Interferometers

For dimensional measurements laser interferometers realize the reference scale for length measurements. Practical measurements also include the interaction between the measurement reference and the measurement object. All the mechanical connections between the interferometer beam splitter, the moving reflector, and the measurement object establish the so-called measurement chain. Every change in these parts during a measurement, for example, due to thermal expansion, bending, or vibrations, causes a length measurement error, which are often larger than the errors of the interferometer itself. For high precision the measurement axes of both systems must have the same direction and must be in line. A parallel arrangement causes a high sensitivity against angular movement, resulting in so-called Abbe errors [101,102]. The laser interferometer can be arranged very flexibly to find optimized solutions for different measurement tasks.

In precision engineering, the main influence results from the temperature [103], because most of the parts to manufacture and to measure are made of metals, often steel. When using an interferometer, it is necessary to measure the refractive index and the temperature of the respective metal part. Especially the temperature measurement causes cost and effort to ensure proper calibration. In fabrication plants the environmental conditions are normally not very stable, so the measurement uncertainty achievable with interferometers suffers from fast variations in the refractive index. Linear encoders can be manufactured with the same coefficient of thermal expansion as steel, so in many practical cases some thermal compensation is realized, and no temperature measurement system is immediately necessary. Those incremental scales have also large time constants, filtering out the influence of short periodic temperature changes, common in production environment. For higher precision grating interferometers with scales made from glass ceramics with a low coefficient of thermal dilatation combined with a temperature measurement system can be used to realize an accuracy comparable with laser interferometers, but at a lower price [104]. Scales are sensitive to the forces of the mounting. Therefore, it is necessary to calibrate machines after assembling the linear encoders. This can be done with defined and reproducible mounted encoders, which have been calibrated with a laser interferometer under stabilized environmental conditions or directly with an interferometer [105]. Especially for large machines the calibration with an interferometer is the only possible solution, because large scales would be difficult to handle and to mount. Laser interferometers are also used for the correct fabrication of grating scales [106].

A typical application of high-precision laser interferometers is the position measurement of two dimensional stages. For quality control in the semiconductor industry, two-dimensional objects, e.g. reticles, must be measured with extremely high repeatability of a few tenth of a nanometre. The principle of a mask measurement system is shown in Figure 2.16. For high-precision 2-D laser interferometric measurements, it is necessary to use plane mirror interferometers with an L-shape mirror. Using a single independent measurement system on stacked axis as widely used in machine tools is too sensitive against varying deviations from the straightness of the moving axis, which are not as stable over time as the L-shape mirror. During motion in one of the axis the flatness deviations of the L-shape mirror cause measurement errors in the other axis. Special measurement set-ups to calibrate the flatness of the mirror have been proposed [107]. As seen in Figure 2.18, a fixed microscope is used to detect structures

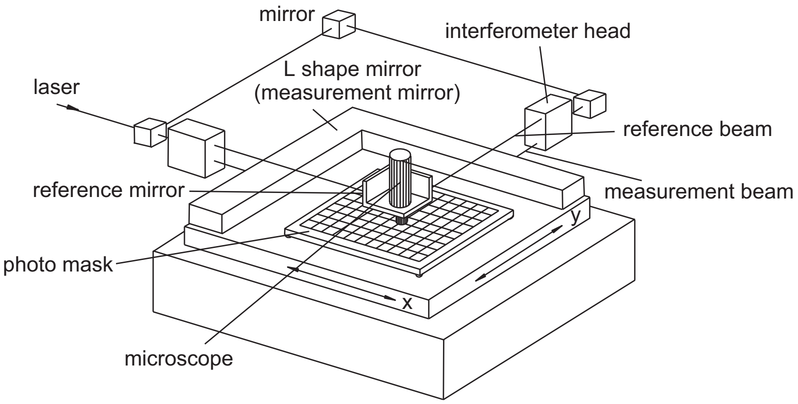

mirror
interferometer head
L shape mirror
(measurement mirror)
laser
reference beam
reference mirror
measurement beam
y
photo mask
x
microscope

FIGURE 2.16 Basic principle of a mask-measuring machine.

on the photo mask. A small L-shape reference mirror is connected to the microscope. The reference beam of the interferometer measures against this mirror. Thus, the interferometer compensates all unwanted motions of the microscope caused, for example, by thermal dilatation of the microscope holders. Today, often additional parallel interferometer axes are used to simultaneously measure the tilt of the slide [28] for online or offline corrections. In wafer scanners it is common to use up to 15 measurement axes to control the position of different moving parts during high-velocity motion.

Another example for high-precision displacement measurements by means of laser interferometry is the detection of gravitational waves. These tiny ripples of space time cause relative length changes in the order of 10^{-21}. Due to this smallness, it took 100 years from their theoretical prediction by A. Einstein in 1915 [108] to the first detection by the LIGO-VIRGO scientific collaboration in 2015 [109]. Figure 2.17 shows the basic layout of the advanced LIGO (Laser Interferometer Gravitational Wave Antenna) interferometer. An Nd:Yag laser with a wavelength of 1064 nm is used. Beyond the classical set-up of a Michelson interferometer [110], the gravitational wave detector additionally incorporates mode cleaners to suppress the propagation of higher order TEM modes, arm cavities, power recycling mirrors, and signal recycling mirrors [111]. The arm cavities locked at resonance enhance the optical path length of the light in the interferometer arms. The power recycling mirror reinjects the light travelling back to the laser and thereby increases the light intensity in the interferometer. Finally, the signal recycling mirror resonantly enhances the gravitational wave signal by sending it back into the interferometer [112].

Due to high-sensitivity requirements, all components of the interferometer, particularly in the arm cavities, need to be carefully designed to mitigate any kind of noise that may arise and perturb the measurement signal. As fundamental noise sources seismic noise [113,114], thermal noise [115–117] and quantum noise have to be circumvented [118,119].

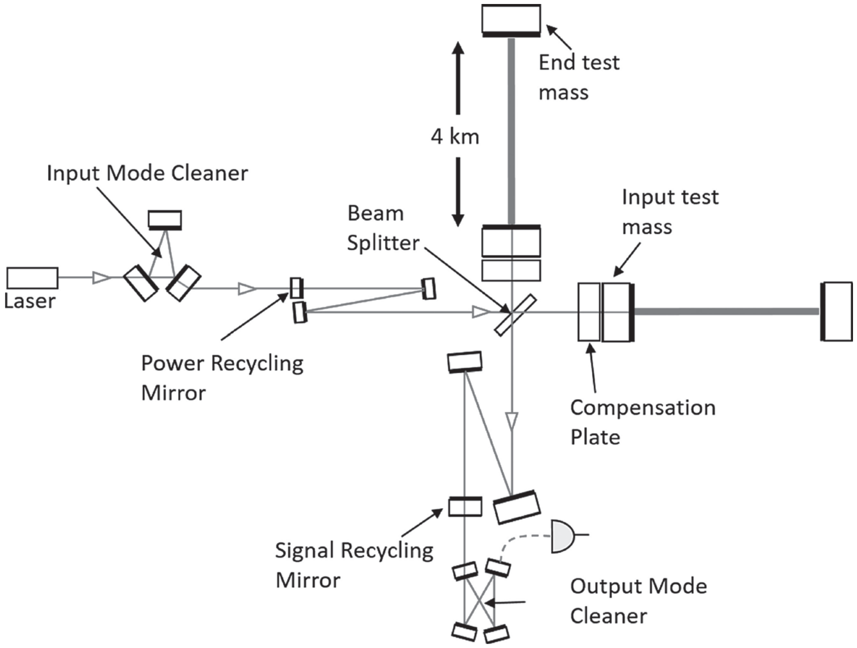

End test
mass
4 km
Input Mode Cleaner
Beam
Splitter
Input test
mass
Laser
Power Recycling
Mirror
Compensation
Plate
Signal Recycling
Mirror
Output Mode
Cleaner

FIGURE 2.17 Basic layout of the LIGO interferometer. The illustration was adapted from [111].

Noise minimization in a gravitational wave detector is a multidimensional optimization problem and requires the consideration and optimization of the optical, mechanical, and thermal properties of the involved components. The noise source, which is directly related to the properties of the laser light, is quantum noise, i.e. the uncorrelated sum of photon shot noise and radiation pressure noise. Shot noise limits the detector's sensitivity at high frequencies, whereas radiation pressure sets a sensitivity limit at low frequencies. The shot noise limit scales with $1/\sqrt{I}$ with the laser intensity I. Thus, to improve the sensitivity limit in terms of shot noise, the intensity of the laser light needs to be increased. Advanced LIGO (aLIGO) uses a laser intensity of 180 W. However, a higher laser power also increases the radiation pressure noise, i.e. the back action on the test masses, whose signal-to-noise ratio scales with $\sqrt{I}$. The trade-off that has to be made between radiation pressure and shot noise is a consequence of Heisenberg's uncertainty principle. The minimum of the uncorrelated sum of both noise contributions is the standard quantum limit (SQL). The SQL can be beaten by using squeezed light [119,120]. The feasible squeezing level is limited by the optical losses of the optical interferometer components.

To achieve a laser power of 180 W, the aLIGO laser system was constructed in three stages: the first stage containing a non-planar ring oscillator (NPRO) generates a beam of 2 W. In the second stage this intensity is amplified to 35 W. The second stage acts as a master laser for the third stage, an injection-locked ring oscillator where the laser intensity is finally increased to 180 W. The laser system also contains pre-stabilization and filtering of the laser frequency, the spatial beam profile, the laser pointing direction as well as fluctuations of the laser power. The frequency reference is provided by a cavity made from monolithic fused silica which is thermally and seismically isolated from external disturbances. The signal from the cavity is fed back to an EOM for phase corrections and to the cavity and temperature controls of the NPRO.

For next-generation gravitational wave detectors, the use of erbium lasers at a wavelength of 1550 nm for an operation at cryogenic is under consideration [121,122]. It should be stressed that solely development of a highly frequency and intensity stable laser at this wavelength is not sufficient. Moreover, a change of the operation wavelength in a gravitational wave detector requires redesign or adaption of several interferometer components [123]. Thereby, the biggest challenge is the mitigation of thermal noise in the optical components and suspension systems [124].

2.5 Multiple Wavelength Interferometry

2.5.1 Gauge Block Calibration

For some applications it is necessary to measure directly the length or dimension of an artefact instead of displacements. An example is the optical calibration of gauge blocks, which are the most commonly used reference standards for industrial length measurements. Multiple wavelength interferometry can also be used in measuring systems for surface profilometry [125].

The interferometer for calibrating gauge blocks often uses the principle of a Michelson interferometer. The measurement beam illuminates the upper surface of the gauge block and the platen. Both interfere with the reference beam as shown in Figure 2.18.

The length L of the gauge block measured with a wavelength λ_i is

$$L = (m_i + j_i)\lambda_i / 2, \tag{2.8}$$

where m_i is the integer value and ϕ_i is the fractional part of the phase difference. The phase difference between gauge block and platen can only be measured without the additional integer value m_i. To determine the value of m_i, the same measurement is carried out with two or three different wavelengths which gives a set of fractional phase differences. Starting with an approximate length of the gauge block, different sets of m_i must be tested whether they match all conditions of the type 2.8 simultaneously. Based on the measurement uncertainty for the fractional phase and the uncertainty of the length of the gauge block, two or three wavelengths are necessary to determine the length with high accuracy. In earlier times, different lines of a spectral lamp have been used. When using lasers, it is today common to stabilize all lasers having different wavelengths of different colours. One laser used is normally the two-mode-stabilized red He-Ne laser. Also, in some interferometers He-Ne lasers at 543 and 612 nm are used. Other laser sources used today are the frequency-doubled Nd:Yag laser at 532 nm, stabilized on an iodine absorption line, and in the infrared spectrum, a diode laser at 780 nm stabilized on a rubidium absorption line.

If the gauge block and the platen are of different material or have different surface roughness, some corrections have to been taken into account. The phase shift due to the reflection can be calculated from the refractive index n and the absorption coefficient k with $\tan(\delta) = 2k/(1 - n^2 - k^2)$ [126]. The influence of the roughness on the effective plane of reflection is known from experimental and theoretical investigations [127,128]. If the roughness is small compared to the wavelength, the effective plane is the mean plane of the surface. Also effects of wringing the gauge block on the plate must be considered [106]. Different techniques have been developed recently to automatically measure the surface topography of the gauge block [129,130]. Interferometry is the adequate measurement technology to realize traceability of gauge blocks to

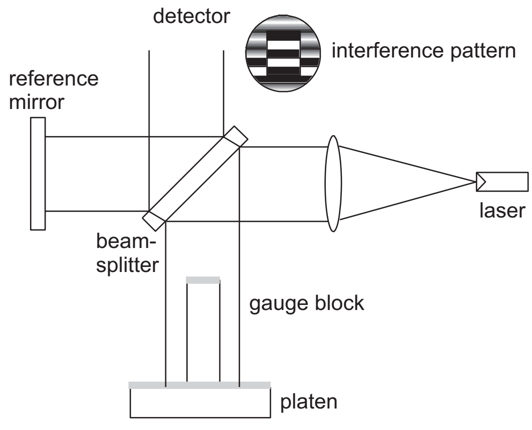


FIGURE 2.18 Principle of interferometric gauge block measurement.

the definition of the unit of length [131]. In industry, calibration of secondary level gauge blocks is mostly done by comparison with interferometrically calibrated gauge blocks using mechanical probing [132].

Gauge block interferometers are also be used for measuring material parameters like the coefficient of thermal expansion, long-time dimensional stability, or compressibility [133].

2.5.2 Interferometric Distance Measurements

For distance measurement of lower accuracy, the efforts for stabilizing multiple laser sources are normally too expensive. An adequate alternative is then the use of a tunable laser [134], a diode laser in particular. When the phase of the interference phase $\phi = m + \varphi$ is measured at two wavelengths the following equations for the distance can be applied.

$$
\begin{aligned}
D &= \phi_1\lambda_1/2 = (m_1+\varphi_1)\lambda_1/2 \\
&= \phi_2\lambda_2/2 = (m_2+\varphi_2)\lambda_2/2 \qquad (2.9)
\end{aligned}
$$

This can be rewritten to

$$
D = (\phi_2 - \phi_1)\frac{\lambda_1\lambda_2}{2(\lambda_1 - \lambda_2)}, \qquad (2.10)
$$

where $\lambda_1\ \lambda_2\ /\ (\lambda_{1\_}\lambda_2)$ is the so-called synthetic wavelength. The intensity measurement of the interference signal again only delivers the fractional part φ. The differences of the integer parts can be determined by counting full signal periods during continuously tuning the laser from λ_1 to λ_2. Mode hopping must not occur in the tuning range. The speed of the measurement object must be slow compared with the tuning frequency of the laser to ensure proper counting. The unambiguous measurement range is half of the synthetic wavelength. For a large measurement range, only a small tuning range is necessary, but as the interpolation now is also based on the synthetic wavelength, a large synthetic wavelength reduces the accuracy. A laser with a large mode-hopping-free tuning range is necessary for high precision. If the synthetic wavelength is small enough that the measuring uncertainty is smaller than $\lambda_i/2$, the full accuracy of the interpolation of a single wavelength could be used. The wavelength λ_1 and λ_2 must be well known.

The most promising laser for distance measurement is the diode laser. But also other lasers, for example, Nd:Yag lasers have been used. Standard LDs have a very short resonator, which reduces the mode-hopping-free range, and cause a low coherence length due to the small reflectivity of the end faces. These disadvantages can be reduced if DFB or DBR diode lasers with integrated grating reflectors are used or if the diode laser has an anti-reflective coating and is coupled to an external resonator as described in 2.2.4.

The second important point of absolute distance interferometry is the accurate determination of the wavelength used for the phase measurements. One solution is the use of a second interferometer with fixed mechanical length as a reference. Any measured change of the interferometer value must be associated with a wavelength change. If the length of the reference is calibrated, the wavelength change can be calculated [135]. For industrial applications this set-up has the advantage, that changes of the refractive index are automatically compensated by the reference. In the same way Fabry-Pérot etalons could be used, if it is necessary to minimize the size of the reference. For a higher accuracy also the well-known absorption lines of iodine or rubidium could be utilized or frequency combs as an more versatile replacement [136].

REFERENCES

1. Quinn, T.J., Mise en pratique of the definition of the metre (2001). *Metrologia*, 2003, **40**, 103–133.
2. J. Guéna, M. Abgrall, D. Rovera, P. Laurent, B. Chupin, M. Lours, G. Santarelli, P. Rosenbusch, M. E. Tobar, R. Li, K. Gibble, A. Clairon, and S. Bize, Progress in atomic fountains at LNE-SYRTE. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, 2012, **59**, 391–410.
3. S. Weyers, V. Gerginov, N. Nemitz, R. Li, and K. Gibble, Distributed cavityphase frequency shifts of the caesium fountain PTB-CSF2. *Metrologia,* 2012, **49**, 82–87.
4. Schnatz, H., et al., First phase-coherent frequency measurement of visible radiation. *Phys. Rev. Lett*., 1996, **76**, 18–21.
5. Bönsch, G., Wavelength ratio of stabilized laser radiation at 3.39 μm and 0.633 mm. *Appl. Opt*., 1983, **22**, 3414–3420.
6. Telle, H., *Optical Frequency Measurements Substantially Simplified*. PTB News, 2000(3).
7. Kovacicch, R.P., U. Sterr, and H. Telle, Short-pulse properties of optical frequency comb generators. *Appl. Opt*., 2000, **39**(24), 4372–4376.
8. Telle, H., et al., Carrier-envelope offset phase control: a novel concept for absolute optical frequency measurement and ultrashort pulse generation. *Appl. Phys. B*, 1999, **69**(4), 327–332.
9. Peck, E.R., Wavelength or length measurement by reversible counting. *J. Opt. Soc. Am*., 1953, **43**(6), 505–509.
10. Rowley, W.R.C., Signal strength in two-beam interferometers with laser illumination. *Opt. Acta*, 1969, **16**, 159–168.
11. Rowley, W.R.C., Some aspects of fringe counting in laser interferometers. *IEEE Trans. Instrum. Meas*., 1966, **IM 15**, 146–149.
12. Downs, M.J. and K.W. Raine, An unmodulated bi-directional fringe counting system for measuring displacement. *Prec. Eng*., 1979, **2**, 85–88.
13. Ilraith, A.H.M., A moiré fringe interpolator of high resolution. *J. Sci. Instrum*., 1964, **41**, 34–37.
14. Jacobs, S., The optical Heterodyn - key to advanced space signaling. *Electronics*, 1963, **12**, 29–31.
15. Dukes, J.N. and G.B. Gordon, A two hundred foot yardstick with graduation every microinch. *HP-J*., 1970, 2–8.
16. Sommargren, G.E., A new laser measurement system for precision metrology. *Prec. Eng*., 1987, **9**, 179–184.
17. Quenelle, R.C. and L.J. Wuerz, A new microcomputer controlled laser dimensional measurement and analysis system. *HP-J*., 1983, 3–13.
18. Balhorn, R., H. Kunzmann, and F. Lebowsky, Frequency stabilization of internal mirror He-Ne lasers. *Appl. Opt*., 1972, **11**, 742–744.
19. Müller, J. and M. Chour, Two-frequency laser interferometric path measuring system for extreme velocities with high accuracy", *Proc. SPIE 2787, Rapid Prototyping*, (26 August 1996); https://doi.org/10.1117/12.248592.

20. Lawall, J. and E. Kessler, Michelson interferometry with 10 pm accuracy. *Rev. Sci. Instrum.*, 2000, **71**, 2669.
21. Weichert, C., P. Köchert, R. Köning, J. Flügge, B. Andreas, U. Kuetgens, A. Yacoot, A heterodyne interferometer with periodic nonlinearities smaller than ±10 pm. *Measure. Sci. Technol.*, 2012, **23**(9), S.94005.
22. Basile, G., A. Bergamin, G. Cavagnero, G. Mana, 93: Phase modulation in high resolution optical interferometry. *Metrologia*, 1991/92, **28**, 455–461.
23. Schwarze, T. S., O. Gerberding, F. G. Cervantes, G. Heinzel, and K. Danzmann, Advanced phasemeter for deep phase modulation Interferometry. *Opt. Exp.,* 2014, **22**, 18214–18223.
24. Nowakowski, B. K., D. T. Smith, and S. T. Smith, Development of a miniature, multichannel, extended fabryperot fiber-optic laser interferometer system for low frequency si-traceable displacement measurement, in *Proceedings of the 29th ASPE Annual Meeting*, 2014.
25. Peck, E.R., Theory of the corner-cube interferometer. *J. Opt. Soc. Am.*, 1948, **43**, 1015–1024.
26. Craig, C.D. and J.C. Rose, Simplified derivation of the properties of the optical center of a corner cube. *Appl. Opt.*, 1970, **4**, 974–975.
27. Bennett, S.J., A double-passed Michelson Interferometer. *Opt. Commun.*, 1972, 428–430.
28. Sommargren, G.E., Linear/angular displacement interferometer for wafer stage metrology. *Proc. SPIE*, 1989, **1088** (Optical/Laser Microlithography II), 268–272.
29. Bobroff, N., Critical alignments in plane mirror interferometry. *Prec. Eng.*, 1993, **15**(1), 33–38.
30. Teimel, A., Technology and applications of grating interferometers in high-precision measurement. *Prec. Eng.*, 1992, **14**, 147–154.
31. Menzel, C. and E. Menzel, Die Beugungserscheinungen optischer Gitter nach Intensität und Phase. *Optik*, 1948, **3**(3), 247–259.
32. Yacoot, A.; Cross, N., Measurement of picometre nonlinearity in an optical grating encoder using x-ray interferometry. *Measure. Sci. Technol.*, 2003, 14(1), S. 148–152.
33. Gerstenkorn, S. and P. Luc, Atlas du Spectre d'Absorption de la Molécule d'Iode. 1978, Laboratoire AIMÉ-COTTON CNRS II, Centre National de la Recherche Scientifique, 15, quai Anatole-France, 75700 Paris.
34. Allan, D.W., Statistics of atomic frequency standards. *Proc. IEEE*, 1966, **54**, 221–231.
35. Baer, T., F.V. Kowalski, and J.L. Hall, Frequency stabilization of a 0.633 mm He-Ne longitudinal Zeeman laser. *Appl. Opt.*, 1980, **19**, 3173–3177.
36. Brand, U., F. Mensing, and J. Helmcke, Polarization properties and frequency stabilization of an internal mirror He-Ne laser emitting at 543.5 nm wavelength. *Appl. Phys. B*, 1989, **48**, 343–350.
37. Niebauer, T.M., et al., Frequency stability measurements on polarization-stabilized He-Ne lasers. *Appl. Opt.*, 1988, **27**, 1285–1289.
38. Hanes, G.R. and C.E. Dahlstrom, Iodine hyperfine structure observed in saturated absorption at 633 nm. *Appl. Phys. Lett.*, 1969, **14**, 362–364.
39. Balling, P., et al., International comparison of $^{127}I_2$-stabilized He-Ne lasers at l = 633 nm using the third and the fifth harmonic locking technique. *IEEE Trans. Instrum. Meas*, 1995, **IM 44**, 173–176.
40. Hu, J., E. Ikonen, and K. Riski, On the nth harmonic locking of the iodine stabilized He-Ne laser. *Opt. Commun.*, 1995, **120**, 65–70.
41. Quinn, T.J., Results of recent international comparisons of national measurement standards carried out by the BIPM. *Metrologia*, 1996, **33**, 271–287.
42. Simonsen, H.R., U. Brand, and F. Riehle, International comparison of two iodine-stabilized He-Ne lasers at l @ 543 nm. *Metrologia*, 1994/95, **31**, 341–347.
43. Ye, J., et al., Absolute frequency atlas of molecular I_2 Lines at 532 nm. *IEEE Trans. Instrum. Meas.*, 1999, **IM 48**, 544–549.
44. Camy, G., C.J. Bordé, and M. Ducloy, Heterodyne saturation spectroscopy through frequency modulation of the saturating beam. *Opt. Commun.*, 1982, **41**, 325–330.
45. Jungner, P.A., et al., Stability and absolute frequency of molecular iodine transitions near 532 nm. *Proc. SPIE*, 1995, **2378**, 22–34.
46. Bjorklund, G.C., Frequency-modulation spectroscopy: a new method for measuring weak absorptions and dispersions. *Opt. Lett.*, 1980, **5**, 15–17.
47. Hall, J.L., et al., Optical heterodyne saturation spectroscopy. *Appl. Phys. Lett.*, 1981, **39**, 680–682.
48. Cordiale, P., G. Galzerano, and H. Schnatz, International comparison of two iodine-stabilised frequency-doubled Nd:YAG lasers at l =532 nm. Accepted for publication in Metrologia, 37(2),177.
49. Macfarlane, G.M., et al., Interferometric frequency measurements of an iodine stabilized Nd:YAG Laser. *IEEE Trans. Instrum. Meas.*, 1999, **IM 48**, 600–603.
50. Hong, F.-L., et al., Frequency comparison of $^{127}I_2$- stabilized Nd:YAG lasers. *IEEE Trans. Instrum. Meas.*, 1999, **IM 48**, 532–536.
51. Kurosu, T., Development of a hybrid laser system: toward an improved working standard at 633 nm. *IEEE Trans. Instrum. Meas.*, 1999, **IM 48**, 550–552.
52. Barwood, G.P., P. Gill, and W.R.C. Rowley, Frequency measurements on optically narrowed Rb-stabilised laser diodes at 780 nm and 795 nm. *Appl. Phys. B*, 1991, **53**, 142–147.
53. Abou-Zeid, A., N. Bader, and G. Prellinger, Rb-stabilized diode laser for dimensional metrology, in *Ultra-Precision in Manufacturing Engineering*, M.W.a.H. Kunzmann, Editor. 1994, Franz Rhiem Verlag: Duisburg. p. 283.
54. Millerioux, Y., et al., Towards an accurate frequency standard at l = 778 nm using a laser diode stabilized on a hyperfine component of the Doppler - free two - photon transitions in rubidium. *Opt. Commun.*, 1994, **108**, 91–96.
55. Velichanskii, V.L., et al., Miniumum line width of an injection laser. *Sov. Tech. Phys. Lett.*, 1978, **4**, 438–439.
56. Fleming, M.W. and A. Mooradian, Spectral characteristics of external-cavity controlled semiconductor lasers. *IEEE J. Quant. Electron.*, 1981, **QE-17**, 44–59.
57. Littman, M. and H. Metcalf, Spectrally narrow pulsed dye laser without beam expander. *Appl. Opt.*, 1978, **17**, 2224.
58. Liu, K. and M. Littman, Novel geometry for single-mode scanning of tunable lasers. *Opt. Lett.*, 1981, **6**, 117–118.
59. Kersten, P., et al., A transportable optical calcium frequency standard. *Appl. Phys. B*, 1999, **68**, 27–38.

60. Drever, R.W.P., et al., Laser phase and frequency stabilization using an optical resonator. *Appl. Phys. B*, 1983, **31**, 97–105.
61. Edwards, C.S., et al., Frequency-stabilized diode lasers in the visible region using Doppler-free iodine spectra. *Opt. Commun*., 1996, **132**, 94–100.
62. Simonsen, H.R., Iodine-stabilized extended cavity diode laser at l = 633 nm. *IEEE Trans. Instrum. Meas*., 1997, **IM 46**, 141–144.
63. Zarka, A., et al., Intracavity iodine cell spectroscopy with an extended-cavity laser diode around 633 nm. *IEEE Trans. Instrum. Meas*., 1997, **IM 46**, 145–148.
64. Ikegami, T., S. Sudo, and Y. Sakai, *Frequency Stabilization of Semiconductor Laser Diodes*. 1995, Boston, MA: Artech House.
65. Donaldson, R.R., Design and construction of the large optics diamond turning machine. *Prec. Eng*., 1989, **11**, 50–51.
66. Flügge, J. and H. Kunzmann, A vacuum interferometer for the comparison of laser interferometers in air and interferometric grating scales, in *VDI-Berichte 1118*, T. Pfeifer, Editor. 1994, VDI.
67. Wilkening, G. The measurement of the refractive index of the air. *Nova Science*. 1988. Commack.
68. Edlén, B., The refractive index of air. *Metrologia*, 1966, **2**, 71–80.
69. P. E. Ciddor. Refractive index of air: new equations for the visible and near infrared. *Appl. Opt.,* 1996, **35**, 1566–1573.
70. Estler, W.T., High accuracy displacement interferometry in air. *Appl. Opt*., 1985, **24**, 808.
71. Muijlwijk, R., Update of the Edlén formulae for refractiv index of air. *Metrologia*, 1988, **25**, 189.
72. Bönsch, G. and E. Potulski, Measurement of the refractive index of air and comparison with modified Edlen's formula. *Metrologia*, 1998, **38**, 133–139.
73. Steinmetz, C.R., Sub-micron position measurement and control on precision machine tools with laser interferometry. *Prec. Eng*., 1990, **12**, 12–24.
74. Terrien, J., An air refractometer for interference length metrology. *Metrologia*, 1965, **1**, 80–83.
75. Downs, M.J. and K.P. Birch, Bi-directional fringe counting interference refractometer. *Prec. Eng*., 1983. **5**(3), 7.
76. Birch, K.P., et al., The Effect of variations in the refractive index of industrial air upon the uncertainty of precision length measurement. *Metrologia*, 1993, **30**, 7–14.
77. Matsumoto, H. and T. Honda, High accuracy length measuring interferometer using the two colour method of compensating for the refractive index of air. *Meas. Sci. Techol*., 1992, **3**, 1084–1086.
78. Ishida, A., Two wavelength displacement measuring interferometer using second-harmonic light to eleminate air-turbulence induced errors. *Jap. J. Appl. Phys*., 1999, **28**(3), 473–475.
79. Lis, S.A., Air turbulence compensated interferometer for IC manufacturing. *Proc. SPIE,* 26 May 1995. **2440***, Optical/Laser Microlithography VIII,* 891 doi: 10.1117/12.209314.
80. Wu, C.M.: Periodic nonlinearity resulting from ghost reflections in heterodyne interferometry. *Opt. Commun*., 2002, **215**(1–3), S. 17–23.
81. Heydemann, P.L.M., Determination and correction of quadrature fringe measurement errors in interferometers. *Appl. Opt*., 1981, **20**, 3382–3384.
82. Birch, K.P., Optical fringe subdivision with nanometric accuracy. *Prec. Eng*., 1990, **12**, 195–198.
83. Downs, M.J., et al., The verification of a polarisation insensitive optical interferometer system with subnanometric capability. *Prec. Eng*., 1995, **17**, 84–88.
84. Quenelle, R.C., Nonlinearity in interferometer measurements. *HP-J*., 1983, 10.
85. Bobroff, N., Residual errors in laser interferometry from air turbulence and nonlinearity. *Appl. Opt*., 1987, **26**, 2676–2682.
86. Sutton, C.M., Nonlinearity in length measurement using heterodyn laser Michelson interferometry. *J. Phys. E: Sci. Instrum*., 1987, **20**, 1290–1292.
87. Freitas, J.M.d. and M.A. Player, Importance of rotational beam alignment in the generation of second harmonic errors in laser heterodyne interferometry. *Meas. Sci. Technol*., 1993, **4**, 1173–1176.
88. Xie, Y. and Y. Wu, Zeeman laser interferometer errors for nm precision measurements. *Appl. Opt*., 1992, **31**, 881–884.
89. Rosenbluth, A.E. and N. Bobroff, Optical sources of nonlinearity in heterodyn interferometers. *Prec. Eng*., 1990, **12**(1), 7–11.
90. Knarren, B. A. W.H., S.J. A. G. Cosijns, H. Haitjema, and P.H.J. Schellekens, Validation of a single fibre-fed heterodyne laser interferometer with nanometre uncertainty. *Prec. Eng.,* 2005, 29(2), S. 229–236.
91. Cosijns, S., Displacement laser interferometry with subnanometer uncertainty; Phd Thesis; TU Eindhoven; 2004; https://research.tue.nl/en/publications/displacement-laser-interferometry-with-sub-nanometre-uncertainty.
92. Hou, W. and G. Wilkening, Investigation and compensation of the nonlinearity of heterodyn interferometers. *Prec. Eng*., 1992, **14**, 91–98.
93. Badami, V.G. and S.R. Patterson, A frequency domain method for the measurement of nonlinearity in heterodyne interferometry. *Prec. Eng*., 2000, **24**(1), S. 41–49.
94. Tanaka, M., T. Yamagami, and K. Nahayama, Linear interpolation of periodic error in a heterodyn laser interferometer at subnanometer levels. *IEEE Trans. Instrum. Meas*., 1989. **IM 38**, 552–554.
95. Oldham, N.M., et al., Electronic limitations in phasemeters for heterodyne interferometry. *Prec. Eng*., 1993, **15**, 173–179.
96. Jacobs S.F., J.N. Bradford, and J.W. Berthold Iii, Ultraprecise measurement of thermal coefficients of expansion. *Appl Opt.* 1970 Nov 1; **9**(11):2477-2480. doi: 10.1364/AO.9.002477. PMID: 20094290.
97. Teague, E.C. Nanometrology. in *Proc. of Scanned Probe Microscopy; STM and beyond*. 1991. Santa Barbara CA.
98. Dorenwendt, K. and G. Bönsch, Über den Einfluß der Beugung auf die interferentielle Längenmessung. *Metrologia*, 1976, **12**, 57–60.
99. Mana, G., Diffraction effects in optical interferometers illuminated by laser sources. *Metrologia*, 1989, **26**, 87–93.
100. Kunzmann, H., Nanometrology at the PTB. *Metrologia*, 1991/92, **28**, 443–453.
101. Bryan, J.B., The abbe principle revised: an updated interpretation. *Prec. Eng*., 1979, **1**, 129–132.
102. Hosoe, S., Laser interferometric system for displacement measurement with high precision. *Nanotechnology*, 1991, 88–95.

103. Bryan, J.B., International status of thermal error research (1990). *Annals CIRP*, 1990, **39**, 645–656.
104. Kunzmann, H., T. Pfeifer, and J. Flügge, Scales vs. Laserinterferometers. Performance and comparison of two measurement systems. *Annals CIRP*, 1993, **42**, 753–767.
105. Baldwin, R.R., Machine tool evaluation by laser interferometer. *HP-J.*, 1970, 12–13.
106. Yoshiike, K. and E. Akabane, 1.5 m master scale exposure system for high precision scale grating. *Mitutoyo Tech. Bull.*, 1989, 20–27.
107. Ebert, E., Flatness measurements of mounted stage mirrors. *Proc. SPIE,* 19 July 1989, **1087**, Integrated Circuit Metrology, Inspection, and Process Control III, 415. doi: 10.1117/12.953114.
108. Einstein, A., Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik*, 1916, **7**, 769–822.
109. Abbott, B.P., et al. (LIGO Scientific Collaboration and Virgo Collaboration), Observation of gravitational waves from a binary black hole merger. *Phys. Rev. Lett*., 2016, **116**, 061102.
110. W.A. Edelstein, J. Hough, J.R. Pugh, and W. Martin, Limits to the measurement of displacement in an interferometric gravitational radiation detector. *J. Phys. E: Sci. Instrumen*., 1978, July, **11**, 710.
111. Harry, G.M., et al., Advanced LIGO: the next generation of gravitational wave detectors. *Classical Quant. Gravity*, 2010, **27**, 084006.
112. Strain, K.A. and Meers, B.J., Experimental demonstration of dual recycling for interferometric gravitational-wave detectors. *Phys. Rev. Lett*., 1991, **66**, 1391.
113. Abbott, R., et al., Seismic isolation for advanced LIGO. *Classical Quant. Gravity*, 2002, **19**, 1591–1597.
114. Cumming, A.V., et al., Design and development of the advanced LIGO monolithic fused silica suspension. *Classical Quant. Gravity*, 2012, **29**, 035003.
115. Harry, G.M., et al., Thermal noise in interferometric gravitational wave detectors due to dielectric optical coatings. *Classical Quant. Gravity*, 2002, **19**, 897.
116. Braginsky, V.B., Yu. Levin, and S.P. Vyatchanin, How to reduce suspension thermal noise in LIGO without improving the Q of the pendulum and violin modes. *Measure. Sci. Technol*., 1999, **10**, 598.
117. Kroker, S., et al., Brownian thermal noise in functional optical surfaces. *Phys. Rev. D*, 2017, **96**, 022002.
118. Bounanno, A. and Y. Chen, Quantum noise in second generation, signal-recycled laser interferometric gravitational-wave detectors. *Phys. Rev. D,* 2001, **64**, 042006.
119. Abadie, J., et al., A gravitational wave observatory operating beyond the quantum shot-noise limit. *Nat. Phys*., 2011, **7** 962.
120. The LIGO Scientific Collaboration, A gravitational wave observatory operating beyond the quantum shot-noise limit: squeezed light in application, arXiv:1109.2295 [quant-ph] 2011.
121. Punturo, M., et al., The Einstein Telescope: a third-generation gravitational wave observatory. *Classical Quant. Gravity*, 2010, **27**, 194002.
122. The Einstein Telescope Science Team, ET design study document, ET-0106C-10, http://www.et-gw.eu/etdsdocument, 2011.
123. Hild, S., et al., Sensitivity studies for third-generation gravitational wave observatories. *Classical Quant. Gravity*, 2011, **28**, 094013.
124. Nawrodt, R., S. Rowan, J. Hough, M. Punturo, F. Ricci, and J.-Y. Vinet, Challenges in thermal noise for 3rd generation of gravitational wave detectors. *Gen. Relativ Gravit*., 2011, **43**, 593–622.
125. Tiziani, H.J., Heterodyne Interferometry using two wavelength for dimensional measurements. *Proc. SPIE*, 1991(Laser Interferometry IV: Computer-Aided Interferometry).
126. Twaite, E.G., Phase correction in the interferometric measurement of end standards. *Metrologia*, 1977, **14**, 53–62.
127. Ishikawa, S., G. Bönsch, and H. Böhme, Phase shifting interferometry with a coupled interferometer: application to optical roughness of gauge blocks. *Optik*, 1992, **91**, 103–108.
128. Ogilvy, J.A., *Theory of Wave Scattering from a Random Rough Surface*. 1991, Bristol: Adam Hilger.
129. Robinson, D. and G. Reid, *Interferogram Analysis. Digital Fringe Measurement Techniques*. 1993, Bristol and Philadelphia: IOP Publishing.
130. Bönsch, G. and H. Böhme, Phase determination of Fizeau interferences by phase shifting interferometry. *Optik*, 1989, **82**, 161–164.
131. Bönsch, G., Gauge blocks as length standards measured by interferometryor comparison: length definition, tracebility chain and limitations. *Proc. SPIE*, 1998. **3477** (Recent developments in optical gauge block metrolopy).
132. Faust, B.S., J.R. Stoup, and E.S. Stanfield, Minimising error sources in gauge block mechanical comparison measurements. *Proc. SPIE*, 1998. **3477** (Recent developments in optical gauge block metrolopy): p. 127–136.
133. T. Middelmann, A. Walkov, and R. Schödel, State-of-the-art cyrogenic CTE measurements of ultra-low thermal expansion materials. *Proc. SPIE*, 2015, **9574**, 95740N. doi:10.1117/12.2187928.
134. Dale, J., B. Hughes, A.J. Lancaster, A.J. Lewis, A.J.H. Reichold, and M.S. Warden, Multi-channel absolute distance measurement system with sub ppm-accuracy and 20 m range using frequency scanning interferometry and gas absorption cells. *Opt. Expr*., 2014, 22(20), 24869.
135. Thiel, J., Pfeifer, T., and Hartmann, M., Interferometric measurement of absolute distances of up to 40 m. *Measurement*, 1995, **16**, 1–6.
136. Y. Salvadé, N. Schuhler, S. Lévêque, and S.L. Floch: High-accuracy absolute distance measurement using frequency comb referenced multiwavelength source. *Appl. Opt*., 2008, **47**, 2715–2720.
137. Riehle, F., P. Gill, F. Arias & L. Robertsson, The CIPM list of recommended frequency standard values: guidelines and procedures. *Metrologia*, 2018, **55**, 188–200.
138. Quinn, T. J. Practical realization of the definition of the metre. *Metrologia*, 1999, **36**, 211–244.
139. Stone, J. A., Decker, J. E., Gill, P., Juncar, P., Lewis, A., Rovera, G. D., and Viliesid, M. Advice from the CCL on the use of unstabilized lasers as standards of wavelength: the helium-neon laser at 633 nm. *Metrologia*, 2009, **46**, 11–18.

3

Laser Velocimetry

Cameron Tropea

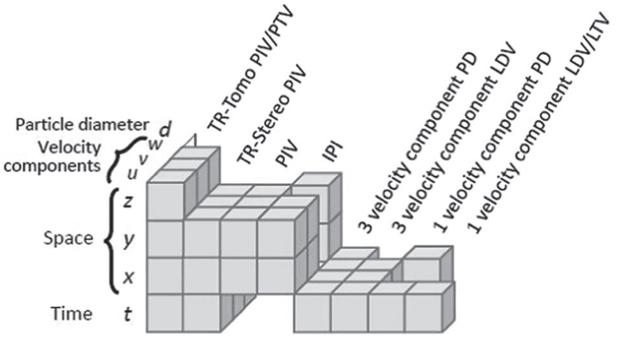
CONTENTS

3.1 Laser Velocimetry 23
    3.1.1 Laser Doppler Velocimetry 24
        3.1.1.1 Fringe Model 25
        3.1.1.2 Doppler Model 25
        3.1.1.3 Transmitting Optics 26
        3.1.1.4 Receiving Optics 27
        3.1.1.5 System Configurations 27
        3.1.1.6 Signal Processing 29
        3.1.1.7 Data Processing 30
3.2 Particle Image Velocimetry 31
    3.2.1 Basic Principles 31
    3.2.2 Choice of Laser 33
    3.2.3 Three-Velocity Component PIV 33
3.3 Doppler Global Velocimetry 34
3.4 Phase Doppler Techniques 35
    3.4.1 Basics of Light Scattering 35
    3.4.2 Measurement Principle 36
    3.4.3 Implementation 38
3.5 Application Issues 38
3.6 Future Directions 39
References 40
Articles 40
Further Reading 42

3.1 Laser Velocimetry

Laser velocimetry encompasses all laser based techniques for measuring one or more velocity components over one or more dimensions. Its most prominent area of application is for the measurement of fluid flow velocity, although the various techniques can easily be applied to the measurement of solid surfaces, dispersed two-phase flows, or the surface of granular flows. Laser velocimetry has the important attributes of being non-intrusive and remote, which has allowed it to displace conventional probe-based techniques in many flow measurement applications. This is particularly true for complex turbulent flows, flows with separation and/or recirculation, flows with small dimensions, or other flows in which probes would either disturb the flow or for which probe access is difficult.

The spectrum of laser velocimetry can be viewed in terms of the number of velocity components measured and in terms of the number of dimensions over which they are measured, as pictured graphically in figure 3.1. The common denominator of these techniques is that they are all tracer-based methods. Particles or tracers are added to the gas or liquid flow to be studied and the velocity of these particles is measured. If the tracer particle is chosen properly in terms of size and density for the particular flow at hand, its velocity can be deemed representative of the



LDV – Laser Doppler Velocimetry
LTV – Laser Transit Velocimetry
PIV – Particle Image Velocimetry
PTV – Particle Tracking Velocimetry
PD – Phase Doppler
IPI – Interferometry Particle Imaging

FIGURE 3.1 Overview of laser velocimetry techniques for fluid flow measurements.

local flow velocity. Applied to two-phase flows, laser velocimetry can capture directly the velocity of the dispersed phase. The issue of tracer particle generation, insertion into the flow and their light scattering properties is often instrumental to the success and accuracy of the measurement, and this issue is elaborated briefly in Sections 3.1 and 3.5. However, the very nature of tracer-based techniques brings with it a random sampling of the flow field, either in space for planar techniques (PIV – Particle Image Velocimetry; DGV – Doppler Global Velocimetry) or in time for point measurement techniques (LDV – Laser Doppler Velocimetry; Laser Transit Velocimetry). This necessitates special considerations with respect to the estimation of flow properties from the sampled data. These aspects are treated together with the signal processing in the final subsection of Section 3.1.

All of the techniques included in Figure 3.1 and discussed in this section rely on elastic light scattering from the tracer particles; although in special applications, fluorescent particles have also been employed. LDV, also widely referred to as laser Doppler anemometry (LDA), was the first of the techniques to be used for flow measurements, with a first instrument being demonstrated by Yeh and Cummins (1964), in which an optical configuration, subsequently known as the reference beam mode, was introduced. First commercial instruments appeared at the beginning of the 1970s. The measurement principles of LDV will be presented in Section 3.1.

Whereas LDV is an interferometric technique, PIV or particle tracking velocimetry (PTV) is an imaging technique. Widespread use of PIV was made possible only with the advent of high-resolution digital recording media (CCD cameras), powerful signal processors, and high-powered pulsed lasers. The PIV technique has undergone rapid development, related to the availability of new hardware components. The extension of the PIV technique to three components and three dimensions was demonstrated first by Elsinga et al. (2006) and is generally known as volumetric (or tomographic) (PIV/PTV). More recently, modified pulsed lasers and fast CCD cameras have become available, allowing time-resolved PIV, up to several thousand frames per second. The PIV/PTV technique will be presented in Section 3.2.

A further planar method, known as either DGV or planar Doppler velocimetry (PDV), utilizes directly the Doppler shift of light scattered from moving particles. This technique is discussed in Section 3.3. As shown in Figure 3.1, laser velocimetry techniques can also be extended to measure particle size (d). This is possible under the assumption of spherical, homogeneous particles and the technique is known as phase Doppler (PD). Techniques for measuring particle size will be discussed in Section 3.4. Further details are given in article D2.68.4 – PD Technique.

3.1.1 Laser Doppler Velocimetry

LDV is an interferometric method of non-intrusively measuring a single-velocity component at a highly localized point in space. In its most widespread application area of fluid flow measurement, one or more optical systems can be combined to extend a system up to three velocity components simultaneously. The high achievable spatial and temporal resolution makes the LDV technique particularly well suited to the measurement of turbulent flow fields.

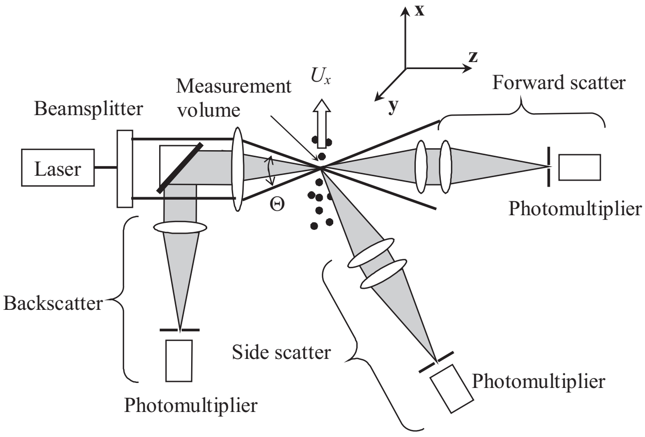


FIGURE 3.2 LDV dual-beam optical arrangement showing different possible collection aperture placements.

The most common optical configuration is the so-called dual-beam arrangement, first introduced by Lehmann (1968) and vom Stein and Pfeifer (1968) and pictured in Figure 3.2. The measurement volume is formed by the intersection of two laser beams, typically coming from the same laser source and focused to a waist of 40–300 μm at the intersection point. Light scattered from tracer particles passing through the volume is collected over an aperture and focused onto a photodetector. The LDV is functional regardless of the collection aperture position, forward scatter, side scatter, or backscatter; however, the scattering properties of the particles must be considered and the forward scatter arrangement is generally preferable, because the scattered light intensity is much higher than in other directions. The signal generated by the photodetector (e.g. shown in Figure 3.12a) is burst-like in nature, increasing in amplitude as a tracer particle enters into the Gaussian intensity profile of the beams and reducing again as the particle leaves. This signal exhibits an inner modulation of frequency

$$f_{\mathrm{D}} = \frac{2U_x \sin(\Theta / 2)}{\lambda} = \frac{U_x}{\Delta x} \tag{3.1}$$

where Θ is the full intersection angle of the beams, λ is the laser wavelength, and U_x is the velocity component of the tracer particle perpendicular to the beams' bisector. The beam intersection and the interference phenomenon giving rise to the Doppler signal are pictured in Figure 3.3. Determining the frequency of each Doppler burst allows the velocity of each individual tracer particle to be determined, since Θ and λ can be known with high accuracy. Statistics over a large number of particles yield then statistics of the flow velocity. Particle rates of several kHz are

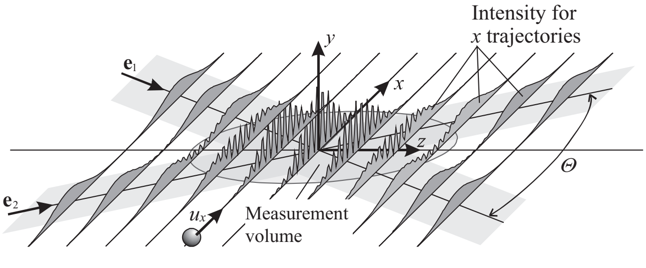


FIGURE 3.3 Intersection volume pictured using the local light intensity.

common, thus achieving a high temporal resolution of the flow velocity fluctuations. It is also evident that the LDV technique requires no calibration, other than possibly the exact measurement of Θ when high precision is required.

3.1.1.1 Fringe Model

The relation (3.1) can be derived in two ways, the first being valid for scattering centres significantly smaller than the measurement volume dimensions. In this case, the particle effectively samples the local light intensity as it passes through the measurement volume. In this intersection volume, where also the beam waist is situated, the assumption of plane wavefronts can be made and the two intersecting beams lead to interference planes, as pictured in Figure 3.3 (and Figure 3.6b). Simple geometric considerations lead to a fringe spacing of

$$\Delta x = \frac{\lambda}{2\sin(\Theta/2)} \tag{3.2}$$

thus resulting in expression (3.1) for a particle traversing the volume with the velocity component U_x. The modulation depth of the generated signal will depend on its trajectory through the measurement volume due to the Gaussian beam intensity profile, as also indicated in Figure 3.3. Multiple particles in the volume simultaneously will lead to additive signals; hence, to a decrease in modulation and to phase noise in the signal, even if the particles are traversing with identical velocities. Thus, it is preferable to adjust the tracer particle concentration and the measurement volume dimensions such that only one particle is present at one time. This is then known as the *single realisation* condition (Buchhave et al. 1979). Assuming a Poisson distribution of particles in space, the probability of more than one particle simultaneously in the measurement volume reduces to 0.5% for the condition $\bar{N} < 0.1$, where $\bar{N}$ is the mean number of particles in the measurement volume (Feller 1971). The mean particle concentration, $\bar{n}$, for this condition is then given by

$$\bar{n} < \frac{\bar{N}}{V_o} = \frac{0.1}{V_o} \tag{3.3}$$

where V_o is the volume of the measurement volume, as given below.

The need for spatial and temporal coherence as well as the requirement for a single wavelength for the LDV technique are evident from this derivation and underline the importance of the laser as a light source.

3.1.1.2 Doppler Model

The fringe model described above is not applicable for larger particles, especially when further characteristics of the signal relating to the particle size (PD technique) must be accounted for. In this case, the interaction of each incident beam with the particle is considered individually, resulting in the Doppler model. Considering the situation in Figure 3.4, the light frequency perceived by a receiver, f_r, after the laser light of frequency f_l is scattered by a moving particle and is shifted due to the Doppler effect

$$f_r = f_l \frac{\left(1 - \dfrac{\mathbf{u}\cdot\mathbf{e}_b}{c}\right)}{\left(1 - \dfrac{\mathbf{u}\cdot\mathbf{e}_{\mathrm{pr}}}{c}\right)} = f_l\left(1 - \frac{\mathbf{u}\cdot\mathbf{e}_b}{c} + \frac{\mathbf{u}\cdot\mathbf{e}_{\mathrm{pr}}}{c}\right) \tag{3.4}$$

where $\mathbf{e}_b$ and $\mathbf{e}_{\mathrm{pr}}$ are the unit vectors of the incident beam and the path between the particle and the receiver, respectively, $\mathbf{u}$ is the vector velocity of the particle, and c is the speed of light. If two laser beams, as shown in Figure 3.5, illuminate the particle, the detector will register two frequencies given by the expressions

$$f_1 = f_l \frac{\left(1 - \dfrac{\mathbf{u}\cdot\mathbf{e}_1}{c}\right)}{\left(1 - \dfrac{\mathbf{u}\cdot\mathbf{e}_{\mathrm{pr}}}{c}\right)} \qquad f_2 = f_l \frac{\left(1 - \dfrac{\mathbf{u}\cdot\mathbf{e}_2}{c}\right)}{\left(1 - \dfrac{\mathbf{u}\cdot\mathbf{e}_{\mathrm{pr}}}{c}\right)} \tag{3.5}$$

where $\mathbf{e}_1$ and $\mathbf{e}_2$ are the unit vectors of each of the incident beams on the particle. Neither of the two frequencies f_1 nor f_2 can be detected directly by the photodetector, since despite the Doppler shift, these are close to the frequency of the incident light. However, the difference frequency (beat frequency) is well within the detection bandwidth of the detector and is equivalent to Equation (3.1) with $f_2 - f_1 = f_{\mathrm{D}}$. This detection of the difference frequency is known as optical heterodyning and is a consequence of the fact that the intensity detected by the photodetector is the square of the sum of all electromagnetic wave amplitudes falling onto the detector surface.

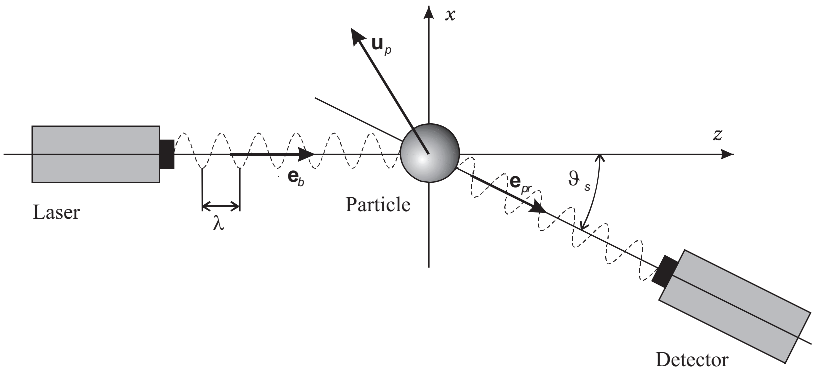


FIGURE 3.4 Doppler shift of light scattered from a moving particle onto a photodetector.

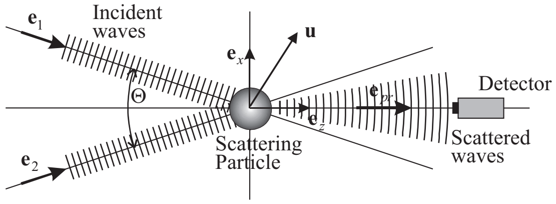

$\mathbf{e}_1$
Incident waves
Θ
$\mathbf{e}_x$
$\mathbf{u}$
$\mathbf{e}_{pr}$
Detector
$\mathbf{e}_z$
Scattering Particle
Scattered waves
$\mathbf{e}_2$

FIGURE 3.5 Scattering of two incident beams from a particle onto a photodetector.

3.1.1.3 Transmitting Optics

The design and layout of any practical laser Doppler system must consider the focussing characteristics of Gaussian beams, since this will determine the dimensions of the measurement volume; hence, the achievable spatial resolution. Generally, a collimator is used to achieve low divergence before the front-focussing lens of the system. The situation pictured in Figure 3.6 can be used to compute the dimensions of the measurement volume. For a focused Gaussian beam, the waist radius after focussing, ω_1, is related to the waist radius before the lens, ω_o by

$$\omega_1 = \frac{\lambda f_\mathrm{T}}{\pi \omega_o} \tag{3.6}$$

where f_T is the focal length of the transmitting lens. The illuminated measurement volume is ellipsoidal in shape, and using Equation (3.6), the half-axes and the volume, V_o, are given by

$$x_m = \frac{\omega_1}{\cos(\Theta/2)} \qquad y_m = \omega_1 \qquad z_m = \frac{\omega_1}{\sin(\Theta/2)} \tag{3.7}$$

$$V_o = \frac{8\pi}{3}\frac{\omega_1^3}{\sin\Theta}$$

For example, a collimated beam of 1 mm diameter from a He-Ne laser ($\lambda = 632.8$ nm) focused using an $f_\mathrm{T} = 310$ mm lens will result in a measurement volume diameter of 250 µm. A beam separation before the lens of $d_s = 40$ mm leads to a full intersection angle of $\Theta = 7.38°$ and a volume of $V_o = 1.3 \times 10^{-10}\,\mathrm{m}^3$.

A further consideration when configuring the transmitting optics is the necessity to sense also velocity direction, since using the principles outlined above, the system cannot distinguish between particles moving in the positive or negative x direction. There are several methods for establishing directional sensitivity; the most commonly used being a frequency shift imposed on one or both of the incident beams. One possibility is to use a Bragg cell as both a beam splitter and a frequency shift device, as shown schematically in Figure 3.7 for a system operating in backscatter.

The frequency shifting of one incident beam relative to the second beam is equivalent to adding a frequency offset f_sh to the Doppler frequency, such that a particle residing in the measurement volume with velocity zero would result in a signal of frequency f_sh. Any motion in the positive direction would increase this frequency and any motion in the negative direction would decrease this frequency. The frequency shift must be chosen larger than the largest expected Doppler frequency in the negative flow direction to avoid any ambiguity. Normally, Bragg cells of 40, 80, or 120 MHZ are employed. These are usually

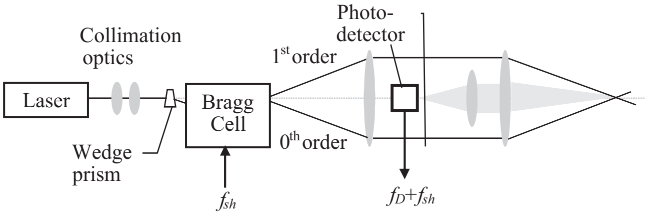

Collimation optics
Photo-detector
1st order
Laser
Bragg Cell
0th order
Wedge prism
f_{sh}
f_D+f_{sh}

FIGURE 3.7 Optical configuration to achieve direction sensitivity using a single Bragg cell functioning as a beam splitter and as a frequency shift device.

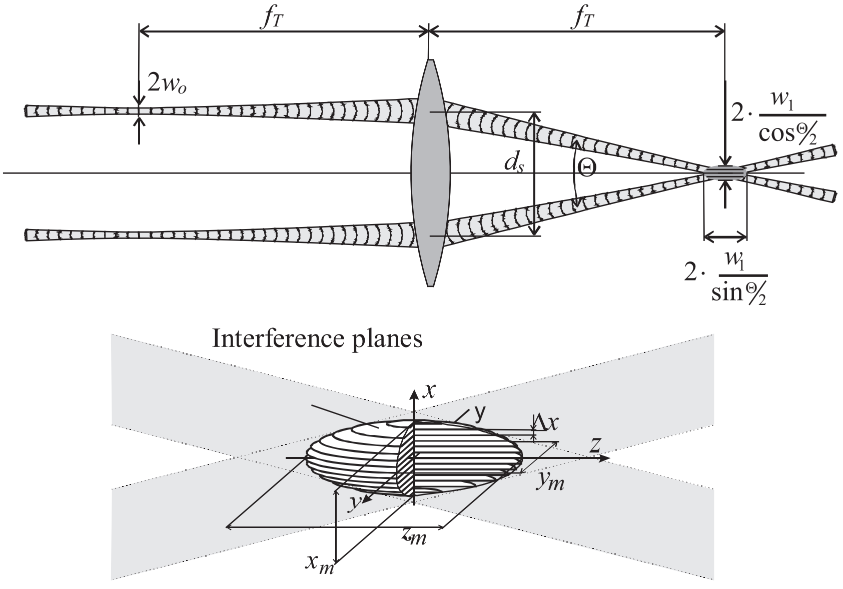

f_T
f_T
$2w_o$
$2 \cdot \frac{w_1}{\cos\Theta/2}$
d_s
Θ
$2 \cdot \frac{w_1}{\sin\Theta/2}$
Interference planes
x
y
Δx
z
y_m
y
z_m
x_m

FIGURE 3.6 Optical dimensions of the measurement volume formed by the intersection of two focused laser beams.

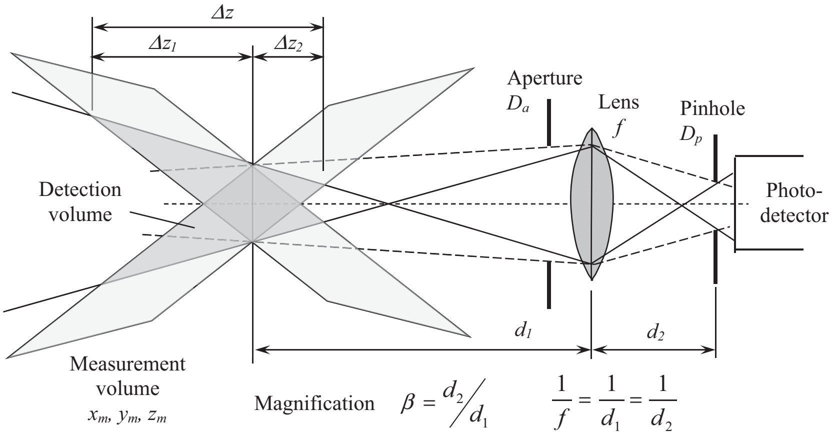


FIGURE 3.8 Dimensioning of the receiving optics for a laser Doppler velocimeter.

implemented in combination with a frequency down-mixing before signal processing, as described further below.

3.1.1.4 Receiving Optics

The receiving optics images the scattered light from the measurement volume onto a photosensitive surface, often a photomultiplier due to their high gain; however, avalanche photodiodes or even PIN diodes can be used, depending on the available power and frequency bandwidth of the scattered light. Such a receiving optics for a simple one-lens system is shown in Figure 3.8. The lens aperture, D_a, and the pinhole aperture, D_p, determine, together with the magnification factor $\beta = d_2 / d_1$, the depth of field and the transverse dimensions of the detected region. For instance, the projection of the pinhole at the focused centre of the measurement volume has the diameter

$$D'_p = {}^{D_p}/_{\beta} \tag{3.8}$$

The effective area from which signals are registered is therefore known as the *detection* volume and must not necessarily coincide with the illuminated *measurement* volume. Especially, a side scatter arrangement can, through the use of a pinhole, greatly reduce the effective length of the detection volume in the z direction. This is important if large flow gradients are to be expected in this direction, since then a spatial averaging of the velocity field will occur, which is generally undesirable. A very common situation for which this technique is used is for the measurement of velocity profiles across a boundary layer, in which gradients are high and dimensions are small. Normally, the system is aligned in the flow such that the largest velocity gradients are along the y coordinate, which also corresponds to the smallest detection volume dimension.

3.1.1.5 System Configurations

A large percentage of commercial LDV systems in use are operated in backscatter because of the convenience of only requiring one optical access to the point of measurement. This is advantageous when further windows would be difficult, e.g. in the cylinder of an internal combustion engine or when the simultaneous traversing of the transmitting optics and receiving optics for acquiring spatial velocity profiles would be difficult, e.g. in large wind tunnels. Nevertheless, there is a price to pay for operating in backscatter or side scatter as opposed to forward scatter, since the scattering intensity of light is much lower. This is illustrated in Figure 3.9, showing a polar diagram of the scattering intensity from a water droplet of various diameters in air.

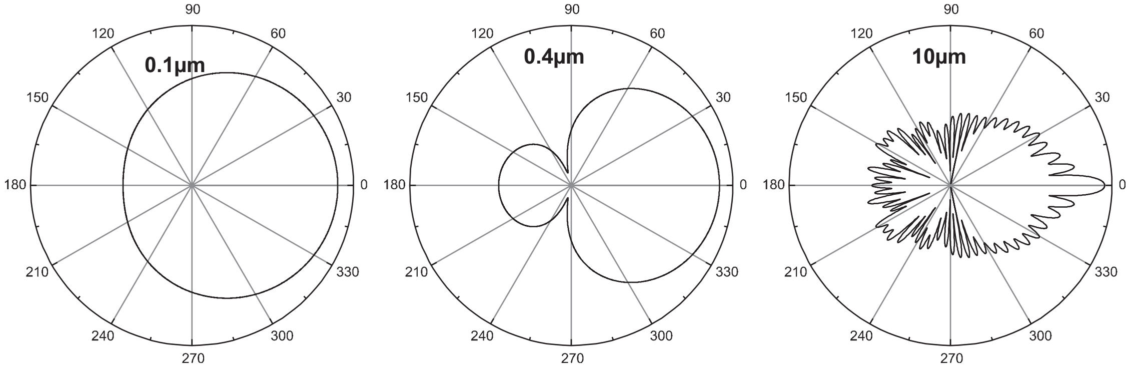


FIGURE 3.9 Polar diagrams of scattered light intensity ($\lambda = 514.5$ nm) from a water droplet in air $m = 1.33$). The amplitude scale is logarithmic.

For typical tracer particle sizes ($d > 1$ µm), the intensity in forward scatter may be 500–1000 higher than in side scatter or backscatter. For all scattering angles, the size dependence of the scattered intensity resembles the curve shown in Figure 3.10, computed also for a water droplet in air at the scattering angles of $\vartheta_s = 0°$ and $\vartheta_s = 90°$. The three diagrams shown in Figure 3.9 correspond to particles in the Rayleigh, Mie, and geometrical optics range, respectively, as marked in Figure 3.10. In laser anemometry, particles in the Mie or geometrical optics range are typically used, where the scattering intensity increases with the square of the diameter. However, increasing particle size indefinitely to improve the scattering amplitude conflicts with the requirement that the particle must also follow all flow fluctuations.

Perhaps, the most common LDV in use today is the two-velocity component, backscatter arrangement, constructed using a fibre-optic link between the transmitting optics/detectors and the measurement head. Such a system is illustrated schematically in Figure 3.11.

In this system, an Ar-Ion laser source has been used and a colour-splitting prism isolates the green ($\lambda = 488.0$ nm) and blue ($\lambda = 514.5$ nm) lines of the laser for coupling into the fibre links. A Bragg cell is used for beam splitting and frequency shifting, as described together with Figure 3.7.

Monomode, polarization-preserving fibres are used to transmit four beams (2 green, 2 blue) to a measurement head, where they are coupled out, collimated, and focused into two overlapping but orthogonally orientated measurement volumes. The measurement head contains also a receiving optics, which images the two measurement volumes onto a larger diameter, graded-index fibre for transmission to the photodetectors. On the receiving side, it is not necessary to preserve coherence or polarization. The scattered light from each volume is divided using a colour separator and interference filter, before reaching

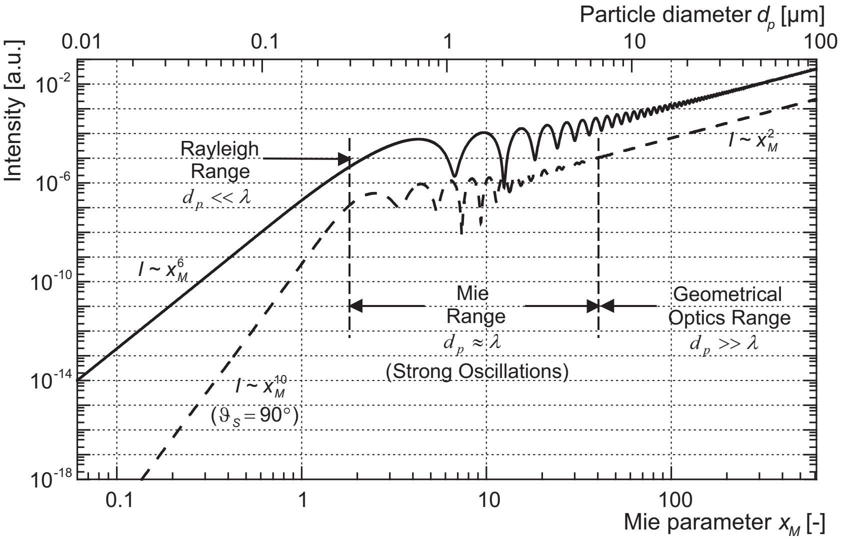


FIGURE 3.10 Scattered light intensity for as a function of the particle diameter for two scattering angles: $\vartheta_s = 0°$ and $\vartheta_s = 90°$

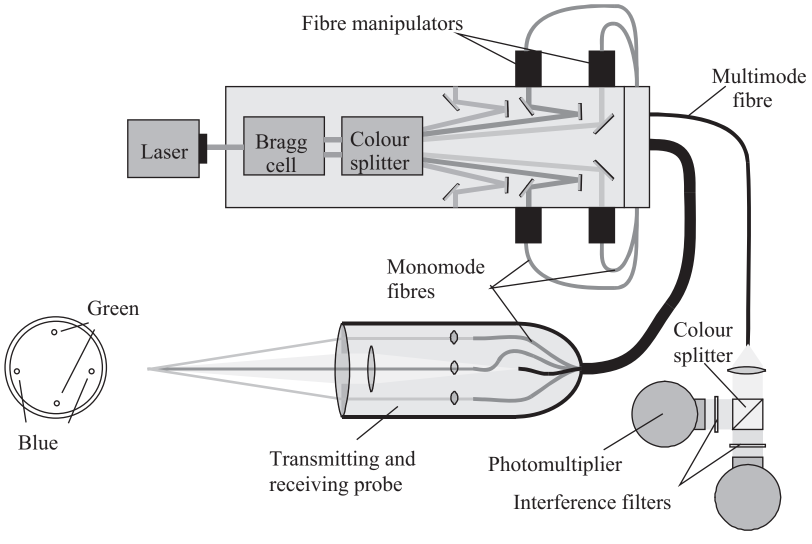


FIGURE 3.11 Two-colour, four-beam laser Doppler system, suitable for measuring two-velocity components. An Ar-Ion gas laser is used as an illumination source.

the photodetector. This system allows the two velocity components perpendicular to the axis of the measurement head to be measured for each particle passing through the volume.

An extension of this system to measure three velocity components is possible by using a third line of the Ar-Ion laser ($\lambda = 476.5$ nm). However, the beam transmission into the measurement volume must be displaced from the two-velocity measurement head to achieve adequate resolution of the third component (Chevrin et al. 1993, James et al. 1997).

Although gas lasers (air-cooled or water-cooled) are the most common light source for LDV systems, these lasers are large and bulky and exhibit low efficiencies of about 0.1%. Solid-state light sources such as continuous-emitting laser diodes or powerful diode-pumped, solid-state lasers are now much more common. Generally, a stable wavelength and a long coherence length are desirable, which exclude many industrial-grade laser sources. Laser diodes, however, exhibit poor beam quality, requiring correction optics, and the power of the fundamental mode is limited to about 200 mW. Nevertheless, this allows a further degree of miniaturization and ruggedness to be achieved, exploited in a number of commercial LDV systems for the measurement of surface velocities. Using diffractive optics, e.g. gratings as beam splitters, LDV systems employing multi-mode laser-diode sources have been demonstrated (Schmidt et al. 1992, Czarske 1999).

Higher power levels are available from diode-pumped solid-state lasers, reaching 300 mW for each wavelength. Furthermore, the wavelengths available are very close to those of Ar-Ion gas lasers; hence, newer laser sources can often replace Ar-Ion lasers with no change of optical components. Details about alternative laser sources for LDV systems can be found in Czarske (2006).

3.1.1.6 Signal Processing

For each particle passing through the detection volume, a burst-like signal is obtained, whose frequency corresponds to one velocity component of the particle, as given by Equation (3.1). Such a signal is shown in Figure 3.12a. Depending on the flow velocity and the particle concentration, a series of such signals occur in time, according to the arrival statistics of the particle in the detection volume. Such a time series is illustrated in Figure 3.12c, after the signal has been passed through a high-pass filter. The signal processor must therefore determine the frequency of each burst, its arrival time, and the burst duration.

There are several approaches to realizing LDV signal processors; however, with high-performance computers now readily available, most systems use a fast analogue-to-digital converter and realize the parameter estimation in firmware and software. Two main algorithms are employed: computation of the power spectral density (PSD) function of signal bursts and estimation of the dominating frequency, and use of the autocorrelation function (ACF) from which a zero-crossing analysis yields the signal modulation frequency. Both approaches have been realized in real-time instruments and provide an automatic adaptation of processed signal length to the input burst length (Lading 1987, Ibrahim and Bachalo 1992, Lading and Andersen 1988, Jensen 1992).

The performance of the signal processor can be evaluated in terms of its expectation and variance of the frequency estimate. Estimators based on the PSD and the ACF are generally bias-free, meaning their expectation exhibits no systematic error. Their variance is limited by the so-called Cramer-Rao Lower Bound (CRLB) (Kendal and Stuart 1963), for Doppler signals given approximately as (Rife and Boorstyn 1974, Wriedt et al. 1989)

$$\sigma_f^2 \geq \frac{3f_s^2}{\pi^2 N(N^2-1)\mathrm{SNR}} \tag{3.9}$$

where N is the number of digitized sample points in the input signal, sampled at a frequency f_s. σ_f represents the lowest possible standard deviation achievable without having *a priori* information about the signal. It clearly represents the lower limit of measurable flow turbulence, since this residual statistical scatter of the estimated signal frequencies can no longer be distinguished from actual flow velocity fluctuations. The scatter arises from the fact that the laser Doppler technique involves several stochastic processes: light scattering, photon detection, and electronic amplification. Most common signal processors presently come very close to achieving the CRLB.

The signal-to-noise ratio (SNR) is an important parameter in determining not only the CRLB but also the actual achieved statistical variance. The SNR is defined as the ratio of signal power (σ_s^2) to noise power (σ_n^2). Noise arises from such sources as shot noise, thermal noise in the electronics, secondary scattering in

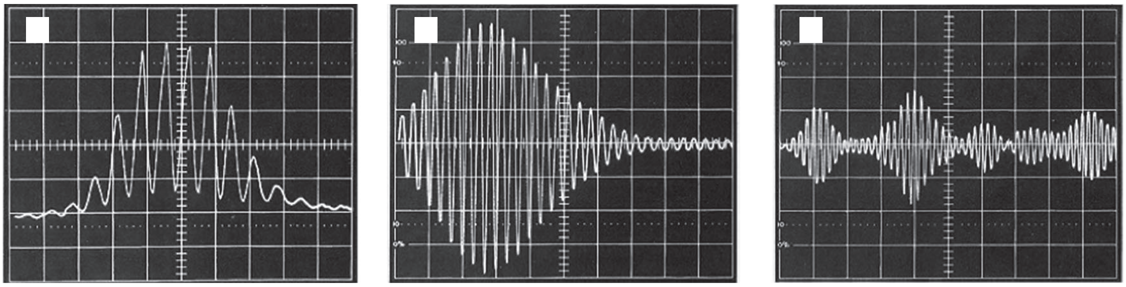


FIGURE 3.12 Illustration of typical LDV signals: (a) signal obtained from the photodetector, (b) signal after passing a high-pass filter, (c) series of filtered Doppler series in time.

the flow system, etc., and is generally considered to be spectrally white. It is proportional to (Stieglmeier and Tropea 1992)

$$\mathrm{SNR} \propto \frac{\eta P_o}{\Delta f}\left(\frac{D_a d_s}{f_{\mathrm{T}} f_{\mathrm{R}}}\right) d^2 \mathrm{GV}^2 \tag{3.10}$$

where η is the quantum efficiency of the detector, P_o is the incident light power before scattering, Δf is the system bandwidth, D_a is the receiving aperture diameter, d_s is the beam separation before focussing, f_{T} and f_{R} are the transmitting and receiving lens focal lengths, d is the particle diameter, G is the Mie scattering function, and V is the modulation depth of the signal. This equation is of immediate use to examine methods in which the SNR of a particular system could be improved: higher laser power or quantum efficiency, dimensioning of the optical system, or choice of scattering particles.

A sample burst signal, its PSD function, and its ACF is shown in Figure 3.13. This signal has a relatively low SNR of 5 dB. Nevertheless, the modulation frequency of the signal is clearly distinguishable in the PSD function and the ACF exhibits a clean oscillation frequency. The noise contribution appears in the PSD as a base level of white noise, constant for all frequencies. In the ACF, the noise, being uncorrelated with itself, appears all in the amplitude of the first coefficient. Both functions are excellent methods of separating noise from signals (Bendat and Piersol 1986). The actual frequency estimation from the PSD function is performed using a curve fit of the function in the immediate neighbourhood of the maximum peak. Typically, the nearest three or five points in the function are used. An interpolation of the frequency at which the fitted curve attains a maximum is then made (Hishida et al. 1989, Matovic and Tropea 1991). A Gaussian curve is generally used for the interpolation, which becomes a parabolic curve when the logarithmic values of the PSD function are considered.

Some additional algorithms and/or electronics are required for signal detection, i.e. to determine what portions of the input signal belong to each burst. More advanced signal detection systems work with an online PSD function and monitor the SNR (Qiu et al. 1994, Bachalo et al. 1989, Jensen 1990). The burst duration is measured as the time over which the *SNR* exceeds some threshold value.

As already pictured in Figure 3.12, a high-pass filter is often employed at the input stage. This removes the low-frequency component (pedestal) of the signal arising from the Gaussian beam intensity distribution across the measurement volume width. A low-pass filter is also often employed to remove high-frequency noise components of the signal. Obviously, the cut-off frequencies of these filters must be carefully chosen to avoid suppressing signal contributions. Many processors now use parallel filter banks and processing, choosing the most appropriate result retroactively in a validation step.

In many cases, the optical system has an imposed frequency shift. Examples have been given above of a Bragg cell functioning also as a beam splitter, but the main reason for frequency shift is to allow directional sensitivity. Therefore, the signal from the photodetector has a frequency equal to the imposed frequency shift plus the Doppler frequency. Since the frequency shift is usually of the order of 40–120 MHz, it is desirable to first reduce this frequency to a range more easily sampled, while maintaining the directional sensitivity. Down-mixing is an electronic mixing of a signal with a stable reference frequency with the input signal, yielding an output signal with the difference frequency of the two signals. This is achieved by means of electronic heterodyning. The principle is easily understood using two sine waves of different frequencies, $s_1 = A_1 \sin\omega_1 t$ and $s_2 = A_2 \sin\omega_2 t$. These signals are added and then squared together to yield a signal of the form

$$\begin{aligned} s = (s_1 + s_2)^2 &= A_1 A_2 (\sin^2\omega_1 t + \sin^2\omega_2 t + 2\sin\omega_1 t \sin\omega_2 t) \\ &= A_1 A_2 \left(\sin^2\omega_1 t + \sin^2\omega_2 t + \sin(\omega_1 + \omega_2)t + \sin(\omega_1 - \omega_2)t\right) \end{aligned} \tag{3.11}$$

If this signal is now passed through a low-pass filter with a cut-off frequency between $\omega_1 - \omega_2$ and ω_1 or ω_2, then only the term with the frequency difference will remain. In laser Doppler systems, ω_1 is the frequency of the reference signal and ω_2 is the Doppler frequency. If the driving frequency of the Bragg cell has been chosen to be, for example, 40MHz, then with a value of $\omega_1 = 2\pi \times 35$ MHz, the down-mixed signal would yield a Doppler frequency above or below 5MHz, depending on whether the flow velocity was positive or negative.

3.1.1.7 Data Processing

The data available from the signal processor consist of frequency, arrival time, and burst duration for each detected particle. The arrival times are approximately randomly distributed, corresponding to a random spatial distribution of particles in the flow. These input data are used to estimate flow parameters, such as the probability density function (PDF) of the flow velocity. The first moment of the PDF is the mean flow velocity, and the normalized second central moment is the turbulence intensity. For these and further flow quantities, it is necessary to also develop appropriate estimators, their goodness again judged according to their expectation and variance. The estimator for the mean flow velocity will be examined first.

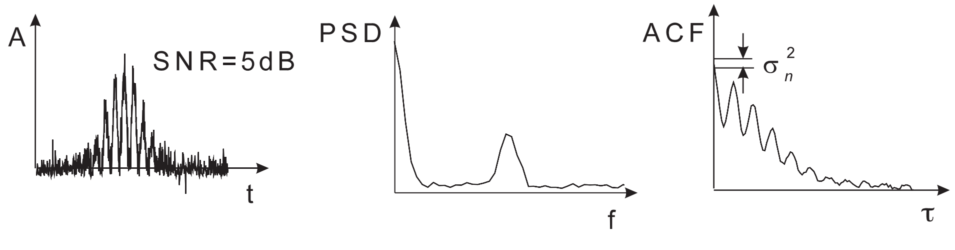


FIGURE 3.13 (a) LDV signal with SNR = 5dB, (b) PSD function of input signal, (c) ACF of input signal.

One peculiarity of LDV data is that the particle arrival rate is usually highly correlated with the measured velocity component. Therefore, during flow fluctuations to higher velocities, more particles will be detected per unit time on average than during flow fluctuations to lower velocities. A simple arithmetic mean of all detected particle velocities will, therefore, yield a positively biased mean estimate, i.e. its expectation is too high. A non-biased mean velocity estimated must account for this particle rate/velocity correlation inherent in the data.

A correct mean velocity estimate can be obtained by using the residence time of each particle in the detection volume, given by the burst length, τ_i (Buchhave et al. 1979).

$$\bar{u}_x = \frac{\sum_{i=1}^{N} \tau_i u_{xi}}{\sum_{i=1}^{N} \tau_i} \tag{3.12}$$

where u_{xi} is the corresponding velocity measured from each burst and N is the total number of detected particles. The residence time is inversely proportional to the absolute velocity and thus to the instantaneous particle rate. This weighted mean estimator is appropriate also for other measured velocity components.

Similar estimators can be written for other flow quantities, for instance, for the variance of velocity fluctuations (Buchhave et al. 1979)

$$\sigma_{u_x}^2 = \frac{\sum_{i=1}^{N} (u_{xi} - \bar{u}_x)^2 \tau_i}{\sum_{i=1}^{N} \tau_i} \tag{3.13}$$

or Reynolds stress terms if a two-velocity component system is being used:

$$\overline{u'_x u'_y} = \frac{\sum_{i=1}^{N} (u_{xi} - \bar{u}_x)(u_{yi} - \bar{u}_y)\tau_i}{\sum_{i=1}^{N} \tau_i} \tag{3.14}$$

The irregular arrival times of the particles in the detection volume also require special consideration when deriving suitable estimators for quantities such as the spectral density of the flow fluctuations. Standard Fast Fourier Transform procedures are not applicable, since these require equidistant sampling of the signal in. Even the Discrete Fourier Transform programmed directly exhibits intolerable variability in its estimate of the spectral density time (Gastor and Roberts 1975; Roberts et al. 1980). There are two solutions to this estimation task. One is the so-called slot correlation; the other is a sample-and-hold procedure with an FIR filter refinement. These techniques are well described in the review article of Benedict et al. (2000).

3.2 Particle Image Velocimetry

PIV and PTV are whole-field techniques of laser velocimetry, measuring one to three velocity components over a two-dimensional plane within the flow. Both PIV and PTV are tracer-based techniques, but neither requires the coherence of laser light, i.e. neither is an interferometric technique (Raffel et al. 2018).

3.2.1 Basic Principles

The principles of operation are shown schematically in Figure 3.14, in which a dual cavity pulsed laser is used to illuminate a plane within the flow twice, in quick succession. The time between pulses must be matched to the local flow velocity; hence, to the expected displacement of the tracer particles between pulses. The pulse duration must be short enough that the motion of the tracer particle is frozen during each exposure, i.e. no streaks on the image. One or two CCD cameras record the scattered light over some defined measurement area, using a separate image for each laser pulse. Such CCD cameras are known as cross-correlation cameras. One camera is sufficient to capture the two, in-plane velocity components. A second camera is necessary if the third, out-of-plane velocity component is to be measured (see below). Any in-plane movement of the detected tracer particles between pulses will then be apparent by comparing the two images. In principle, the PIV/PTV technique can also be realised with a continuous wave (CW) laser. In this case, the pulse duration and time between pulses can be replaced by control of the camera shutter. The disadvantage of this approach is that CW light sources are generally much lower in peak power and therefore are really only applicable for low-speed water flows, where large tracer particles with good scattering properties can be used.

The evaluation of the recorded images to obtain local flow velocity information is performed over small sub-areas (or interrogation areas) of the recorded plane. Within these localized areas, the assumption is made that all particles move with the same flow velocity. Therefore, a cross-correlation of the interrogation area on the two images, together with the magnification of the images, will yield the bulk velocity of the detected ensemble of particles. A velocity vector is obtained for each examined interrogation spot. For a typical CCD camera resolution of 1000 × 1000 pixels, up to 4000 interrogation areas resulting in the same number of velocity vectors per image pair can be obtained. This approach is known as PIV. If individual particle movements are examined – possible only for more sparsely seeded flows or for dispersed two-phase flows – the technique is known as PTV.

If a cross-correlation camera is not available, the two illuminations can be recorded as two or more exposures on the same image. In this case, an autocorrelation is used to find the bulk particle displacement in each interrogation area. Experience shows, however, that this technique is much more susceptible to noise-induced errors in estimating the velocity vectors. An example double-exposure image, a zoom of one interrogation area, and the resulting spatial ACF of that interrogation area are shown in Figure 3.15. The large peak in the ACF at the

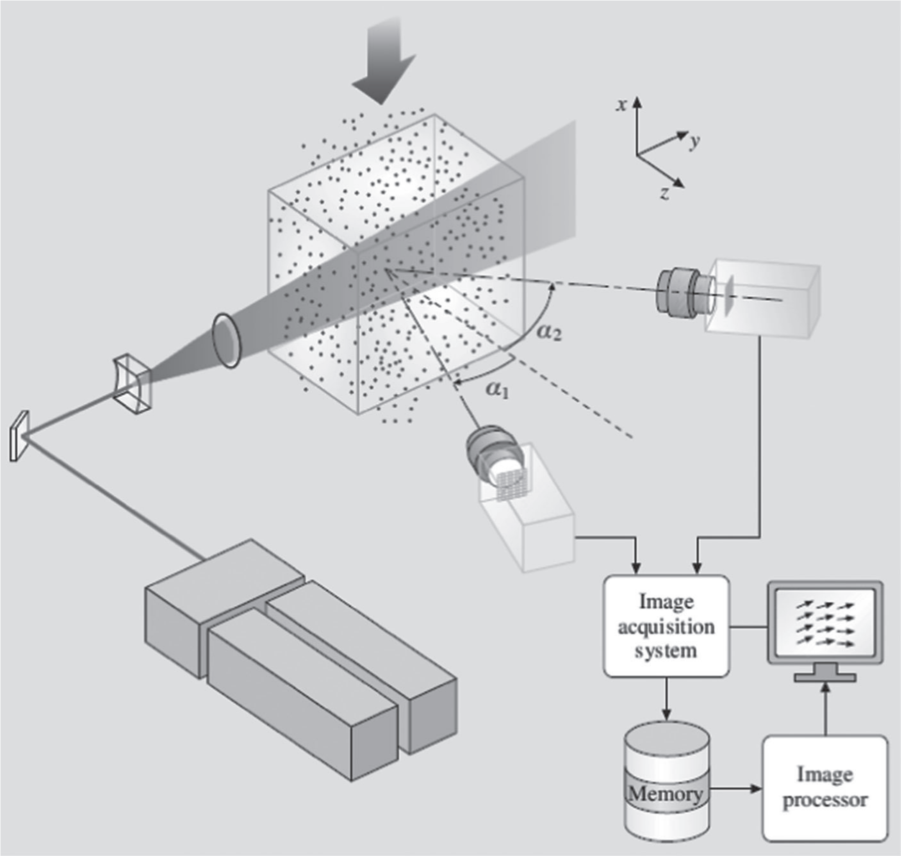


FIGURE 3.14 Schematic representation of a PIV/PTV optical system (Tropea et al. 2007).

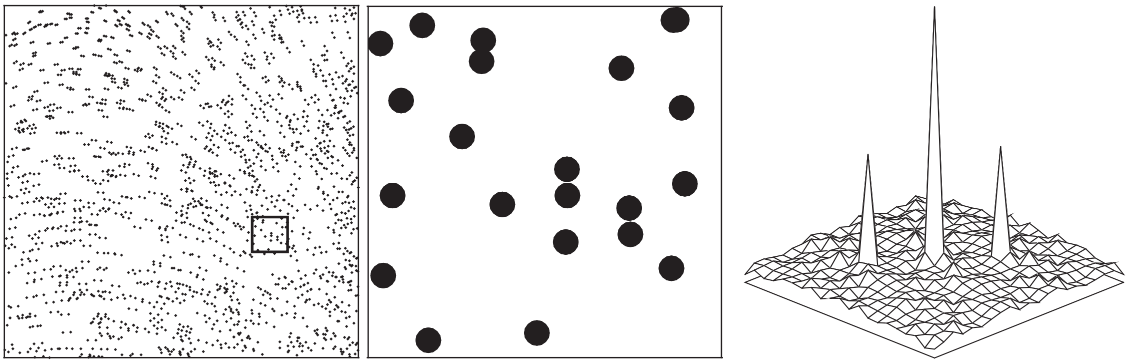


FIGURE 3.15 (a) Double-exposure image, (b) interrogation area **c** spatial ACF.

origin corresponds to the self-products and includes any noise contributions. The two, symmetrically placed smaller peaks correspond to the bulk displacement of the particle ensemble in the interrogation area. From this function, the direction of motion cannot be unambiguously determined. The use of a cross-correlation CCD camera resolves this difficulty.

More refined algorithms for estimating the particle movement within the interrogation area have been introduced, which also take into account local velocity gradients, the so-called local field corrections. These techniques, well reviewed in Scarano (2002), also yield a much higher spatial resolution of velocity vector estimation (Hart 2000, Nogueira et al. 1999.)

An example result using a two-velocity component PIV system is shown in Figure 3.16, showing the leading-edge vortex separating from a pitching plate at a Reynolds number of 60 000. In this presentation of results, a velocity vector has been assigned to each interrogation area. Further processing can examine spatial velocity gradients, e.g. vorticity, strain rate, etc., or spatial means. In Figure 3.16, the out-of-plane-normalized vorticity is shown as a colour shading. Such result images can be obtained in rapid succession, depending on the frequency of the double-pulsed laser, typically about 1000 times per second. Time mean averages of the flow field can then be obtained by averaging individual vectors over a large number of result images.

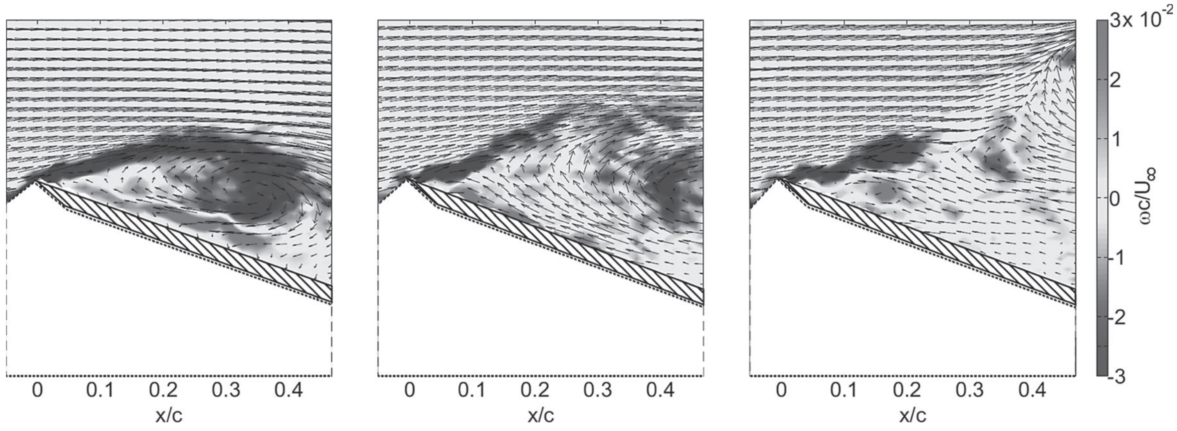


FIGURE 3.16 Example result of a TR-PIV measurement using cross-correlation cameras. (Images taken from pitching plate. Reynolds number 60 000, based on mean velocity. Normalized vorticity shown as colour shading.)

3.2.2 Choice of Laser

Historically, a large number of different lasers have been used throughout the development of the PIV technique. For reasons mentioned above, CW lasers are not attractive any more. The copper-vapour laser ($\lambda = 510\text{nm}, \lambda = 578\text{nm}$) offers a high average power but relatively low pulse energy (10 mJ), whereas the ruby lasers ($\lambda = 694\text{nm}$) deliver very high pulse energy but low repetition rates. The most widely used laser for PIV is the Nd:YAG (neodymium:yttrium-aluminium-garnet) laser, operated with a frequency-doubling KDP crystal ($\lambda = 532\text{nm}$) and at power levels of 40–150 mJ per pulse. Furthermore, a Pockels cell is used as a quality switch (Q-switch) in the cavity to allow the laser to be operated in a triggered mode. Pulse energies of up to several hundred mJ at repetition rates of 10–20 Hz are common.

Time resolution can be achieved by using lasers with higher repetition rates. Typically, dual-cavity, diode-pumped Nd:YLF lasers are used, with pulse energies up to 30 mJ and repetition rates of up to 10 kHz. These laser exhibit excellent beam profiles, allowing thin and homogeneous light sheets to be generated. Special illumination optics are required for volumetric velocimetry, typically achieving height-to-width ratios of 1:1 to 1:10.

3.2.3 Three-Velocity Component PIV

The PIV/PTV technique can be extended to three-velocity components in two dimensions by adding a second CCD camera to the system (Hinsch 1995). The fundamentals of such a stereoscopic system are outlined in Figure 3.17. The true three-dimensional displacement of the particles between exposures $(\Delta x, \Delta y, \Delta z)$ is estimated from a pair of two-dimensional displacements $(\Delta x, \Delta y)$ as seen from left and right cameras. For this configuration, the light sheet must be sufficiently thick so that a majority of the particles remain illuminated for both laser pulses.

Since the cameras are looking at the light sheet plane from an angle, the image would not normally be in focus over the entire area. This is resolved by using the so-called Scheimpflug configuration, in which the recording plane is no longer parallel to the plane of the focussing lens. This arrangement is shown diagrammatically in Figure 3.18. The image plane, lens plane, and object planes for each of the cameras intersect in a common line. Nevertheless, a strong distortion is introduced with this technique, i.e. the magnification is not constant across the image. This difficulty is resolved by first developing a model, describing how the object plane is mapped onto the image plane.

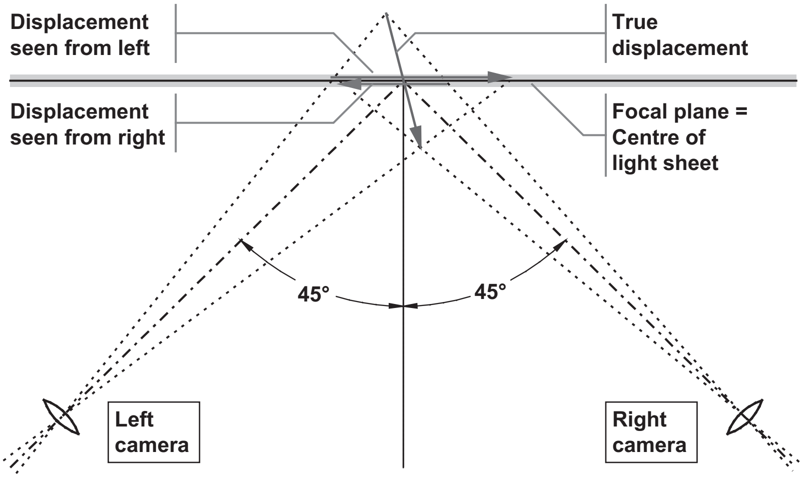

Displacement seen from left
True displacement
Displacement seen from right
Focal plane = Centre of light sheet
45°
45°
Left camera
Right camera

FIGURE 3.17 Optical arrangement of a three-velocity component PIV/PTV system.

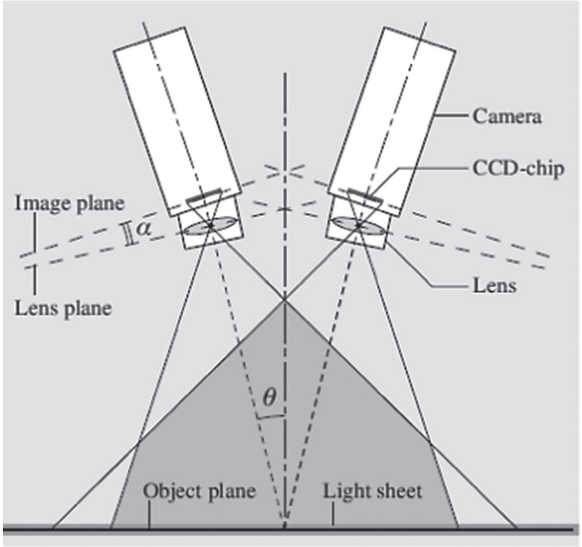


FIGURE 3.18 Scheimpflug arrangement of the CCD cameras for stereoscopic PIV/PTV (Tropea et al. 2007).

The parameters of this model are then determined through a camera calibration. For this purpose, a calibration target, consisting of a regular grid pattern or pattern of markers displaced also in the out-of-plane direction, is placed at the illumination plane. Preliminary images of this calibration target are taken and used to compute the calibration constants (Westerweel and Nieuwstadt 1991, Willert 1997, Raffel et al. 2018).

One example of such a three-component PIV measurement is pictured in Figure 3.19. The velocity vector maps for each of the left and right cameras are shown, together with the derived three-component velocity information. The out-of-plane velocity component is shown in Figure 2.2.19c as shaded contours.

3.3 Doppler Global Velocimetry

A further planar method utilizing the Doppler shift of light scattered from a moving particle is known as DGV or PDV. The Doppler shift of light scattered from a moving particle, given by expression (3.5), can be expressed as (for beam 1)

$$\Delta f_1 = f_l - f_1 = f_l \frac{\left(\mathbf{e}_{\mathrm{pr}}\text{-}\mathbf{e}_1\right)}{c}\,\mathbf{u} \tag{3.15}$$

and has the physical interpretation illustrated by Figure 3.20.

If Δf_1 can be measured and $\mathbf{e}_1$ and $\mathbf{e}_{\mathrm{pr}}$ are known geometrical quantities from the optical arrangement, the component of the particle velocity u along the vector $\left(\mathbf{e}_{\mathrm{pr}} - \mathbf{e}_1\right)$ can be determined. Observing the same particle from three different directions would, in principle, allow all three velocity components of the particle to be determined. As mentioned in Section 3.1, Δf_1 cannot be directly detected electronically by a photodetector and this difficulty was circumvented in the LDV technique by using two illuminating beams and optical heterodyning. However, there exist several methods of optically determining Δf_1. One involves an interferometer coupled to the detection optics, and such an instrument has been realized as a three-velocity component interferometer (Smeets and George 1981).

More widespread is the use of iodine absorption cells, used as a frequency-to-transmission converter (Komine 1990). A schematic representation of one such absorption line is shown in Figure 3.21.

Δf_1 can be determined by measuring the iodine cell transmission of the scattered light compared to the non-attenuated light. The practical realization of this measurement requires a detector before and after the iodine cell to register the net absorption. The laser must be stabilized so that the absolute position on the calibrated absorption line is constant. Furthermore, CCD cameras can be used as detectors, yielding

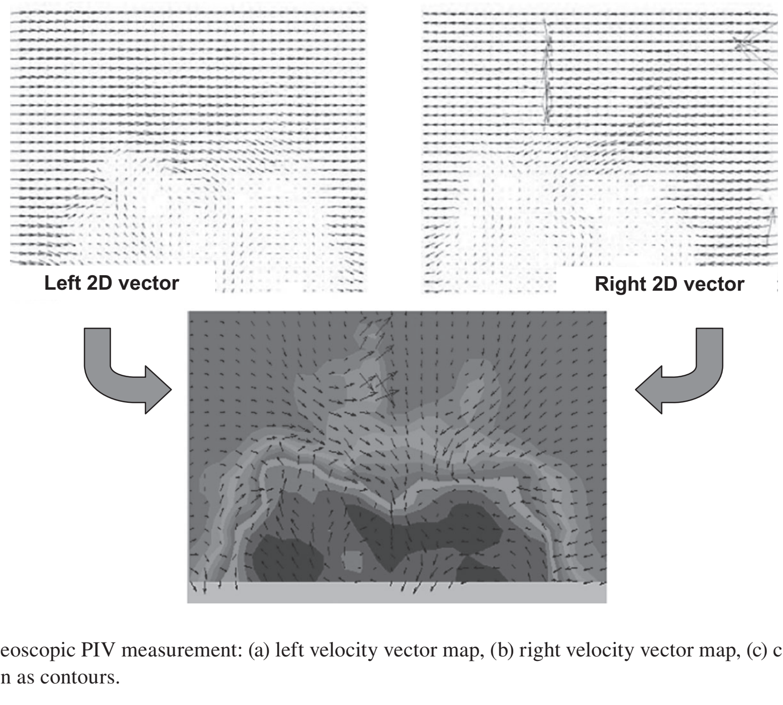


FIGURE 3.19 Example stereoscopic PIV measurement: (a) left velocity vector map, (b) right velocity vector map, (c) composite velocity vector map. Out-of-plane velocity is shown as contours.

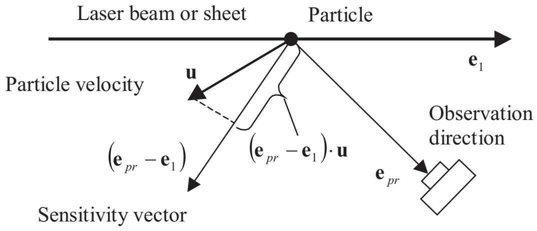


FIGURE 3.20 Sensitivity of Doppler shift to particle velocity.

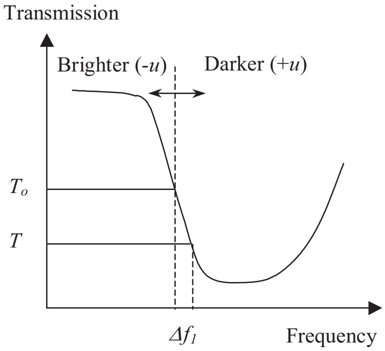


FIGURE 3.21 Schematic representation of the absorption line of an iodine cell.

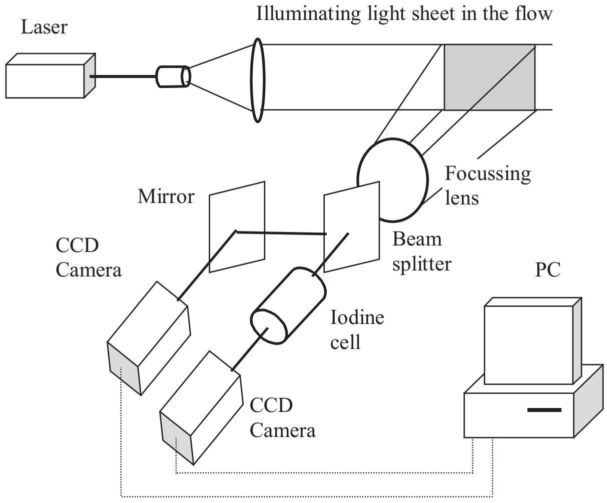


FIGURE 3.22 Doppler Global optical system.

velocity information over an entire plane rather than at a single point. Either CW or pulsed lasers can be used, the latter requiring an additional triggering electronic for the detectors. Depending on the type of laser (CW or pulsed), the result is either time-averaged or instantaneous flow velocities. A typical set-up using a CW laser is illustrated in Figure 3.22.

The extension of the system to two- or three-velocity components can be achieved by adding further synchronized camera systems at various observation directions. Alternatively, several light sheets from different directions can be used. The pictures of each light sheet must be taken one after the other, with the consequence that this method is restricted to stationary flows only. On the other hand, this method is lesser expensive and easier to handle. If long exposure times are used on the cameras, then turbulent velocity fluctuations are integrated on each pixel of the camera chip, yielding a mean velocity. This also allows sparser seeding, at least for stationary flow fields.

The accuracy and resolution of the DGV technique depend on a number of factors, including frequency stability of the laser, calibration and temperature stability of the iodine cell, sensitivity of the cameras, and optical alignment and image de-warping procedures (Meyers et al. 2001, Roehle 1999). One typical analysis performed for a DGV system used to measure a free jet with flow velocities in the range 40–130m/s concluded that the measurement error was ±0.7m/s, although higher values are more typical (Morrison and Gaharan 2001). For this reason, the DGV technique is typically applied to higher speed flows, e.g. for aerodynamics studies (Beutner et al. 2001), turbomachinery components (Roehle et al. 2000), combustion chambers, or supersonic flows (Crafton et al. 2001).

A similar technique to DGV, but based on Rayleigh scattering, was introduced in 1981 by Miles. The Filtered Rayleigh scattering technique also provides measurement of temperature and pressure of a flow, but is also significantly more complex and up to now has been restricted to laboratory experiments (Miles and Lempert 1997, Miles et al. 2001).

3.4 Phase Doppler Techniques

The PD technique is an extension of the laser Doppler technique to allow not only the velocity but also the size of the detected particle to be measured. It is assumed that the particle is spherical and homogeneous and that the relative refractive index of the particle, *m*, being the ratio of the particle refractive index to that of the surrounding medium, is known. The technique is applicable to all relative refractive indexes, e.g. bubbles, particles, or droplets.

3.4.1 Basics of Light Scattering

Before developing the measurement principles of the PD technique, some additional remarks will be made about light scattering from small, spherical particles. The polar scattering plot shown in Figure 3.9c can be unwrapped as shown in Figure 3.23, for each of parallel and perpendicular polarized light. Not only is the total scattered light amplitude shown, as in Figure 3.9, but also the contributions of different scattering modes, including diffraction, reflection, and the first five orders of refracted light. This division of contributions can be interpreted in terms of geometrical optics, as shown in Figure 3.24 for $m > 1$ (e.g. water in air, $m = 1.33$), $m < 1$ (e.g. air bubble in water, $m = 0.75$), and for the general case. The fundamentals of light scattering can be found in textbooks (Van de Hulst 1981, Bohren und Huffmann 2008) and developments that are more recent are reviewed in Wriedt et al. (1998) or Albrecht et al. (2003).

From Figure 3.23, it is important to note that at certain scattering angles, one scattering mode dominates, for instance between 30° and 80°, first-order refraction dominates for perpendicularly polarized light. Reflection dominates near 90° for parallel-polarized light. The intensity relations of different scattering orders may change with particle size and refractive index, and Figure 3.23 is only one typical example.

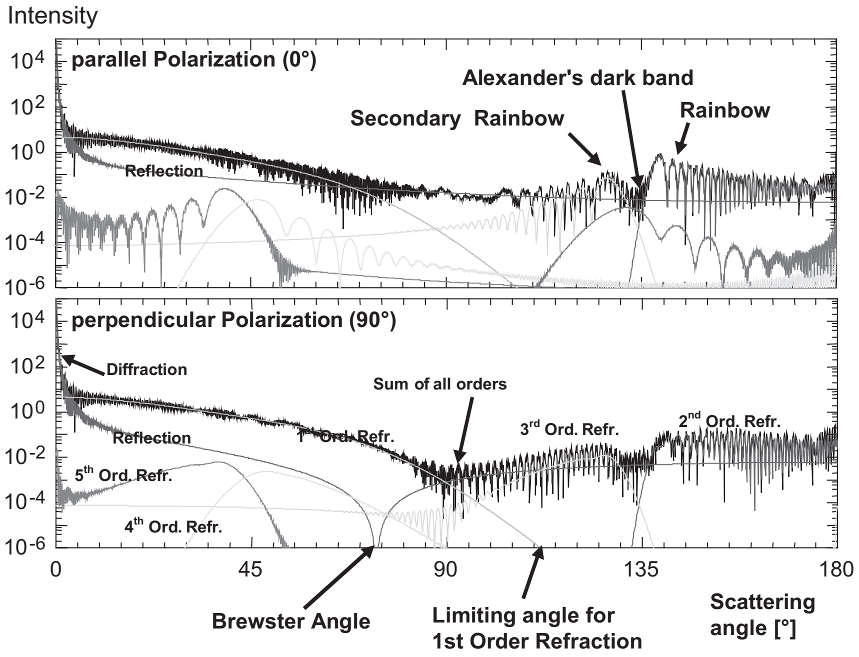


FIGURE 3.23 Scattered light intensity from a spherical particle as a function of scattering angle ($m = 1.33, \lambda = 514.5\,\text{nm}, d = 100\,\mu\text{m}$): (a) parallel polarization and (b) perpendicular polarization.

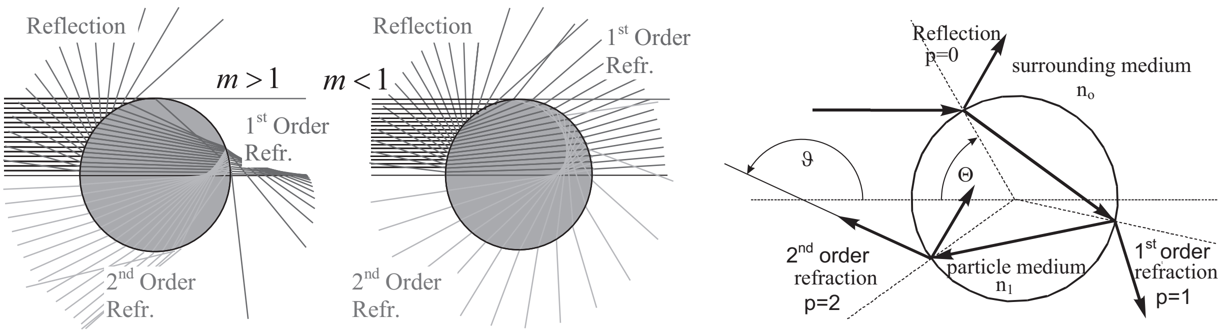


FIGURE 3.24 Representation of scattered light in terms of geometrical optics: (a) $m > 1$, (b) $m < 1$, (c) nomenclature for the general case.

3.4.2 Measurement Principle

The PD technique uses a transmitting optics identical to the laser Doppler technique – two beams intersecting at an angle Θ. However, the detectors of the PD system are placed at a scattering angle, such that one scattering mode dominates the collected light intensity. For the preceding example of water in air, a detector position of 30° off-axis (from forward scatter) would collect primarily light due to first-order refraction.

The operating principle of the PD can now be understood by using path-length difference arguments. This is illustrated in Figure 3.25, showing two in-coming beams, which arrive at the detector after first-order refractive scattering. The points at which the beams enter the particle are known as incident points and the points at which the beams leave the particle are called glare points. For a given particle size, relative refractive index, and detector position, these points are unique for any given scattering order. The points are in fact finite in size due to the finite size of the receiving aperture.

Due to the different path lengths of the beams through the particle, interference fringes appear in the far-field. The interference fringe spacing is directly proportional to the diameter of the particle. As the particle moves through the incident beams, the fringe pattern traverses across the detector and a signal resembling a laser Doppler signal is obtained. The modulation frequency of this signal corresponds to the particle velocity, exactly as given in Equation (3.1) for LDV. However, the fringe spacing remains unknown. For this reason, at least two detectors are required to realize the PD technique and usually three are employed, as shown schematically in Figure 3.26.

Each detector sees the moving interference pattern and produces a similar signal; however, due to their spatial displacement, the signals are phase-shifted. Knowing the frequency and phase shift between the two signals, the fringe spacing and thus the particle diameter can be determined. The diameter can be expressed directly in terms of the measured phase difference, ΔΦ, for instance, for reflected light (Albrecht et al. 2003).

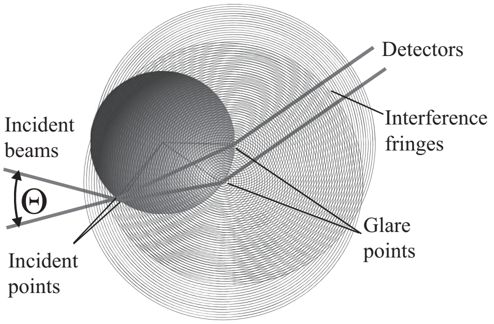


FIGURE 3.25 Path-length difference interpretation of the PD technique.

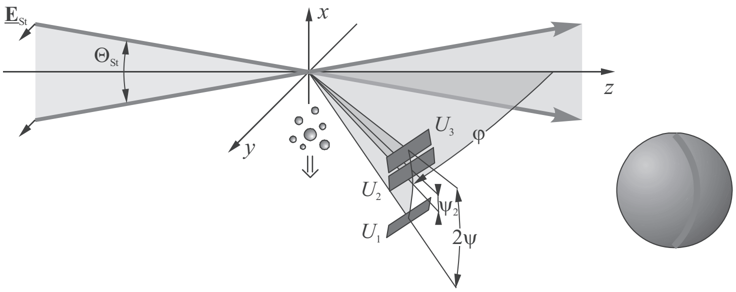


FIGURE 3.26 Schematic representation of detector aperture configuration in a three-detector, standard PD system.

$$\Delta\Phi_{13} = \frac{2\pi d}{\lambda}\sqrt{2}\left(\begin{array}{l}\sqrt{1-\cos\psi\cos\phi\cos{}^{\Theta}\!/_{2}+\sin\psi\sin{}^{\Theta}\!/_{2}}\;-\\ \sqrt{1-\cos\psi\cos\phi\cos{}^{\Theta}\!/_{2}-\sin\psi\sin{}^{\Theta}\!/_{2}}\end{array}\right) \tag{3.16}$$

where ϕ is the off-axis angle of the detectors and $\pm\psi$ are the elevation angles of the detectors 1 and 3 out of the y-z plane. For first-order refraction, the relation reads

$$\Delta\Phi_{13} = \frac{-4\pi d}{\lambda} \times\left(\begin{array}{l}\sqrt{1+m^2-m\sqrt{2}\sqrt{1+\cos\psi\cos\phi\cos{}^{\Theta}\!/_{2}+\sin\psi\sin{}^{\Theta}\!/_{2}}}\;-\\ \sqrt{1+m^2-m\sqrt{2}\sqrt{1\cos\psi\cos\phi\cos{}^{\Theta}\!/_{2}-\sin\psi\sin{}^{\Theta}\!/_{2}}}\end{array}\right) \tag{3.17}$$

Note that with refracted light, the relative refractive index m must be known; however in both cases, the relation between phase difference and particle diameter is linear and no calibration is required. In practice, non-linearities arise for small particles ($< 20\,\mu$m) because other scattering orders become influential. Nevertheless, in many cases, particles down to about 1–2 μm can be measured with a resolution of about ±1μm.

With two detectors, measurements are limited to sizes resulting in phase differences less than 2π – beyond this limit, size ambiguity arises. Employing three detectors, as shown in Figure 3.26, solves this difficulty. The respective phase difference/size relations for the two detector pairs $U_1 - U_3$ and $U_2 - U_3$ are shown in Figure 3.27. The phase difference $\Delta\Phi_{13}$ corresponds to several different particle diameters; the correct choice is indicated by the particle diameter resulting from the $U_2 - U_3$ phase difference, $\Delta\Phi_{23}$. The size agreement between these two diameters can be used as a validation criterion.

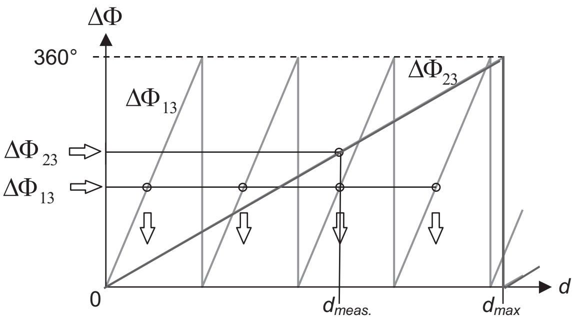


FIGURE 3.27 Phase difference/size relations for a three-detector, standard PD system.

3.4.3 Implementation

As with laser Doppler systems, the PD technique is often implemented using fibre optics. An example PD receiver is shown in Figure 3.28, in which three detecting apertures have been incorporated. The size range is adjusted through choice of the intersection angle Θ, the off-axis angle ϕ, and the elevation angles of the detection apertures $\pm\Psi$. The latter is chosen by selecting different interchangeable aperture masks – two examples are pictured in Figure 3.28. A slit aperture is used on the input face of the receiving fibres. Its projection back onto the illuminated volume effectively limits the length of the detection volume. This is necessary for computation of particle fluxes, which requires a reference area.

The detector aperture size and elevation angle can be varied by selecting different aperture masks. The receiver is linked to the photodetectors using graded-index optical fibres.

The signal processing is similar to LDV in that frequency, arrival time, and burst duration for each particle are determined. Additionally, the phase difference between pairs of signals must be found. This can be achieved either through a cross-correlation or using the cross-spectral density function (Domnick et al. 1988).

The data processing for PD measurements is, however, considerably more complex than for LDV. When compiling statistics about the measured particle sizes, the dependence of the detection volume size on the size of particle must be considered. Larger particles scatter more light; hence, their detection volume is larger than for small particles. Particle statistics must therefore compensate for the fact that more large particles will be detected than small particles. To do this, the effective detection volume size for each particle size class is estimated from the velocity and the maximum signal duration found in each size class. This dimension is used as an inverse weighting for all statistical distributions involving size (Saffman et al. 1984, Qiu and Sommerfeld 1992).

Furthermore, the projected detection volume area changes according to the trajectory of the particle through the volume. This must be accounted for in computing fluxes. For determining the trajectory, a second or even a third velocity component is found to be advantageous (Roisman and Tropea 2002). These velocity components can be added to the system as with LDV systems.

3.5 Application Issues

The choice of technique for measuring flow velocities and/or particle size will depend on many factors, but primarily on the required flow quantities and on their necessary spatial and temporal resolution. The required spatial and temporal resolution is generally determined by the Reynolds number of the flow

$$\mathrm{Re} = \frac{UL\rho}{\eta} \tag{3.18}$$

With U and L being representative length and velocity scales, ρ is the density of the medium and η its dynamic viscosity. The smallest scales of fluid motion in a turbulent flow are given by the Kolmogorov length scale (Tennekes and Lumley 1972, Pope 2000)

$$\eta_K = L\,\mathrm{Re}^{-3/4} \tag{3.19}$$

yielding corresponding time scales of η_K / U for convected turbulence. These scales define the necessary seeding concentration and measurement volume size, if all scales of turbulence are to be resolved. If not all scales of turbulence must be captured, the necessary spatial resolution is often dictated by local mean velocity gradients, e.g. in a boundary layer. For strong gradients in an LDV volume, corrections for turbulence quantities have been derived (Durst et al. 1995, Durst et al. 1998). Similar considerations have been made for accommodating velocity gradients in the processing of PIV data (Scarano 2002).

These timescale estimates of flow velocity fluctuations also dictate the demands on tracer particles to ensure slip-free movement. An exact equation of motion for particles in a flow is rather complex (Crowe et al. 1998), but for the present discussion, the response of tracer particles to flow velocity changes can be considered to be a first-order system with a time constant (relaxation time) of

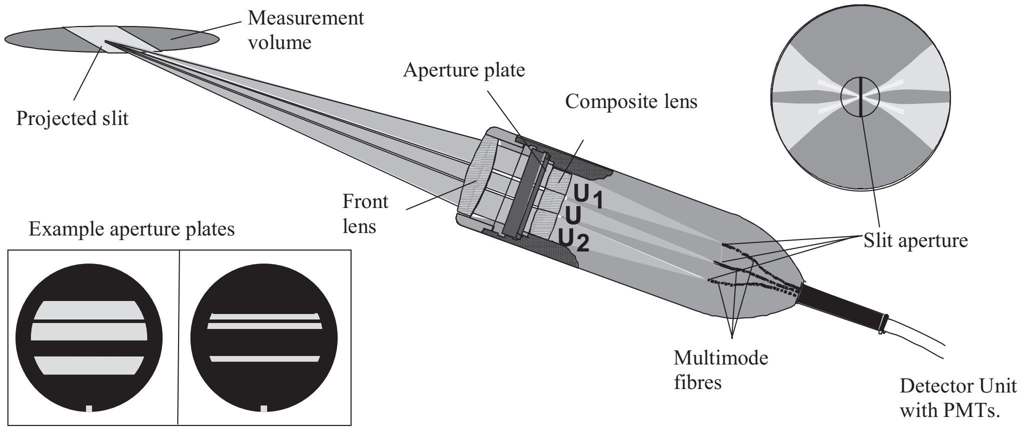


FIGURE 3.28 Example receiving unit of a PD system.

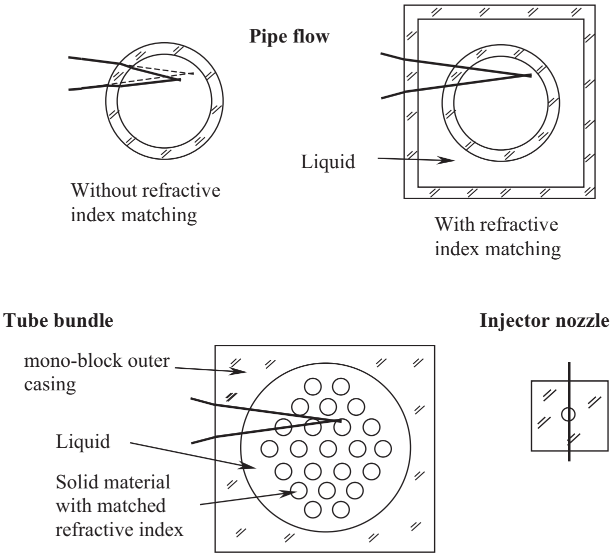

Pipe flow
Without refractive index matching
Liquid
With refractive index matching
Tube bundle
mono-block outer casing
Liquid
Solid material with matched refractive index
Injector nozzle

FIGURE 3.29 Example flow containment systems with refractive index matching.

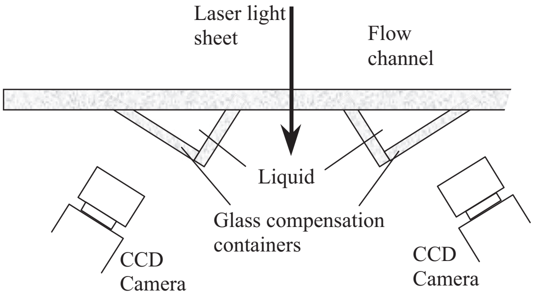


FIGURE 3.30 Optical arrangement of cameras for stereoscopic PIV in a liquid flow.

$$\tau_o = \frac{\rho_p d_p^2}{18\eta} \tag{3.20}$$

where ρ_p and d_p are the density and diameter of the particle, respectively (Melling 1997). A low value of d_p ensures that the tracer particles follow the flow velocity fluctuations closely, even up to high frequencies, suggesting the use of small particles. However, small particles lead to lower scattered light intensity and low SNR values (Equation 3.9). Therefore, special tracer particles are often employed, exhibiting either low densities or high values of the Mie scattering function G, e.g. titanium dioxide (Bang Laboratories). Methods of particle generation and introduction into the flow are reviewed in Melling (1997) and in Albrecht et al. (2003).

All measurement techniques discussed in this section require some degree of optical access to the flow region of interest. This is a very application-specific requirement, which cannot be generalized in a short discussion; however, for liquid flows, some additional considerations arise due to refraction at the containment walls. Entering a liquid vessel with light beams or sheets at non-normal angles can lead to system astigmatism and beam displacement, such that, for instance, the two beams of an LDV system no longer cross (Zhang and Eisele 1995, Zhang and Eisele 1996). This effect becomes even more complex for curved containment walls, e.g. pipe flow (Bicen 1982, Doukelis et al. 1996).

One partial remedy is to use optical windows and working fluid of the same refractive index. An outer containment can then be used to optically remove curved surfaces of containment, as illustrated in Figure 3.29. A good overview of possible solutions is also given in Bai and Katz (2014). For stereoscopic PIV systems, compensation vessels can be used, as shown in Figure 3.30. Some common combinations of liquids and transparent materials for refractive index matching are listed in Table 3.1. In some cases, a mixture of two fluids is used, in which the exact refractive index is modified through the mixing ratio.

3.6 Future Directions

Whereas the LDV technique has reached a high level of maturity and only minor developments can be expected in the future due to technological advances, the field of PIV/PTV is still advancing rapidly in several directions. With the advent of time-resolved measurements of all three velocity components in the three coordinate dimensions, numerous new application areas become feasible. One such advancement exploits the

TABLE 3.1

Some Liquid and Wall Material Combinations for Refractive Index Matching

Containment		Fluid(s)					
Name	n_c [–]	**Name**	n_f [–]	T [°C]	$\nu \times 10^6$	**Laser**	**Reference**
Glass FK-3 (Schott)	1.464	Sohio's mineral seal oil, Sohio's MDI-57		31	6.79		Dybbs and Edwards (1987)
Dura n-50		Diesel oil mixture		20	4.3	He-Ne	Durst et al. (1995)
Pyrex Glass (Corning)	1.474	Dow Corning 556 Dow Corning 550		28.1	43.3		Dybbs and Edwards (1987)
Acryl Glass	1.491	Dow Corning 550 Union Carbide L42		22	188		Dybbs and Edwards (1987)
Acryl Glass	1.491	Tetrachlorethylene 1,1,2 trichlortrifluorethan	1.51 1.36	25	0.524	He-Ne	Herr and Pröbstle (1987)
Acryl Glass	1.491	Diesel oil mixture	1.497 1.475	22.5	8.0	Ar^+	Stieglmeier et al. (1989)

momentum equation of fluid mechanics in which forces on a body can be related to changes in the velocity field; hence, knowing the changes in velocity field from PIV measurements therefore allows one to compute forces on a body. For instance, Ragni et al. (2009) reviews how surface pressure and aerodynamic loads on transonic airfoils can be determined based on PIV. This work has its origins in work from Unal et al. (1998) and is restricted to two-dimensional steady flows. Similarly, the pressure field can be estimated from the measured velocity field (van Oudheusden, 2013). Extensions of these techniques to time-resolved measurements have been demonstrated for both pressure (Neeteson et al. 2016) and for aerodynamic loads in unsteady flows (Kurtulus et al. 2006) and three-dimensional flows (van de Meerendonk et al. 2016). These advances represent vast new possibilities for the study of unsteady flows, in particular, flows in which the model is moving and for which forces and moments are otherwise very difficult to measure using conventional means, such as a force balance (due to overriding inertial forces due to model movement).

A second notable development is the improvement of particle-tracking algorithms to achieve measurements in flows with much high-seeding densities, hence improving the spatial and temporal resolution of turbulence measurements. These algorithms have been termed "Shake-the-Box" and have now been extended to four dimensions, i.e. time-resolved measurements in three coordinate directions. The current state-of-the-art is well reviewed in Schröder et al. (2015).

REFERENCES

ARTICLES

Bachalo WD, Werthimer D, Raffanri R, Hermes RJ (1989) A high speed Doppler signal processor for frequency and phase measurements, *3rd Int. Conf. on Laser Anemom. – Adv. and Appl*., Swansea: late paper.

Bai K, Katz J (2014) On the refractive index of sodium iodide solutions for index matching in PIV. *Exp. in Fluids* 55: 1704

Bang Laboratories, Inc, 9025 Technology Drive, Fishers, IN 46038-2886, USA: http//www.bangslabs.com.

Bendat JS, Piersol AG (1986) *Random Data: Analysis and Measurement Procedures*, John Wiley and Sons, New York.

Beutner TJ, Gregory SE, Williams GW, Baust HD, Crafton J, Campbell DC (2001) Forebody and leading edge vortex measurements using planar Doppler velocimetry. *Meas. Sci. Technol*. 12: 378–394.

Bicen AF (1982) Refraction correction for LDA measurements in flows with curved optical boundaries, *TSI Quarterly*, VIII (2):10–12.

Bohren CF, Huftmann DR (2008) *Adsorption and Scattering of Light by Small Particles,* John Wiley & Sons, Hoboken, NJ.

Chevrin PA, Petrie HL, Deutsch S (1993) The accuracy of a three-component laser Doppler velocimeter system using a single-lens approach, *ASME J Fluids Eng.* 115: 142–147.

Crafton J, Campbell DC, Gregory SE (2001) Three-component phase-averaged velocity measurements of an optically perturbed supersonic jet using multi-component planar Doppler velocimetry. *Meas. Sci. Technol*. 12: 409–419.

Czarske J (1999) Nutzung diodengepumpter Faserlaser in der Laser-Doppler-Anemometrie, Techn. *Messen* 66: 363–371.

Czarske J (2006) Laser Doppler velocimetry using powerful solid-state light sources *Meas. Sci. Technol*. 17: R71–R91.

Domnick J, Ertel H, Tropea C (1988) Processing of phase/Doppler signals using the cross-spectral density function, *Proc. 4th Int. Symp. on Appl. of Laser Techn. to Fluid Mech.,* Lisbon, Portugal: paper 3.8.

Doukelis A, Founti M, Mathioudakis K, Papailiou K (1996) Evaluation of beam refraction effects in a 3D laser Doppler anemometry system for turbomachinery applications. *Meas. Sci. Technol*. 7: 922–931.

Durst F, Jovanovic J, Sender J (1995) LDA measurements in the near-wall region of a turbulent pipe flow, *J Fluid Mech*. 295: 305–335.

Durst F, Fischer M, Jovanovic J, Kikura H (1998) Methods to set up and investigate low Reynolds number, fully developed turbulent plane channel flows, *ASME J Fluid Eng.* 120: 496–503.

Dybbs A, Edwards RV (1987) Refractive index matching for difficult situations, *2nd Int. Conf. on Laser Anemom. – Adv. and Appl*., Strathclyde: paper IP1.

Elsinga GE, Scarano F, Wieneke B, van Oudheusden BW (2006) Tomographic particle image velocimetry. *Exp. in Fluids* 41: 933–947.

Feller W (1971) *An Introduction to Probability Theory and its Applications* Vol I, Wiley, New York.

Gaster M, Roberts JB (1975) Spectral analysis of randomly sampled signals, *J Inst. Maths. Appl.* 15: 195–216.

Hart DP (2000) DPIV error correction, *Exp. in Fluids* 29: 13–22.

Herr W, Pröbstle G (1987) Index matched flow measurements in rod bundles near grid spacers with swirl generators, *2nd Int. Conf. on Laser Anemom. – Adv. and Appl.*, Strathclyde: 117–129.

Hinsch KD (1995) Three-dimensional particle velocimetry. *Meas. Sci. Technol.* 6: 742–753.

Hishida K, Kobashi K, Maeda M (1989) Improvement of LDA/PDA using a digital signal processing systems processor (DSP), *3rd Int. Conf. on Laser Anemom. – Adv. and Appl.*, Swansea, UK: paper S2.

Ibrahim KM, Bachalo WD (1992) The significance of the Fourier analysis in signal detection and processing in laser Doppler and phase Doppler applications, *Proc. 6th Int. Symp. on Appl. of Laser Techn. to Fluid Mech.*, Lisbon, Portugal: paper 21.5.

James SW, Tatam RP, Elder RL (1997) Design considerations for a 3D laser Doppler velocimeter for turbomachinery applications, *Rev. of Sci. Instr.* 68: 3241–3246.

Jensen LM (1990) Coherent frequency burst detector apparatus and method, US Patent No 4,973,969.

Jensen LM (1992) LDV Digital signal processor based on autocorrelation, *Proc. 6th Int. Symp. on Appl. of Laser Techn. to Fluid Mech.*, Lisbon, Portugal: paper 21.4.

Kendal M, Stuart A (1963) *The Advanced Theory of Statistics*, Vol 2. Charles Griffen, London.

Komine H (1990) System for measuring velocity field of fluid flow utilising a laser-Doppler spectral image converter, US Patent 4,919,536.

Kurtulus DF, Scarano F, David L (2006) Unsteady aerodynamics force estimation on a square cylinder by TR-PIV. *Exp. in Fluids* 42: 185–196.

Lading L (1987) Spectrum analysis of LDA signals, *Proc. Int. Specialists Meeting on the Use of Computers in Laser Velocimetry*, ISL, France: paper 20.

Lading L, Andersen K (1988) A covariance processor for velocity and size measurements. *Proc. 4th Int. Symp. on Appl. of Laser Anemom. to Fluid Mech.*, Lisbon, Portugal: paper 4.8.

Lehmann B (1968) Geschwindigkeitsmessung mit dem Laser-Dopplerverfahren, Wiss. Berichte der Telefunken AG 141.

Matovic D, Tropea C (1991) Spectral peak interpolation with application to LDA signal processing. *Meas. Sci. Technol.* 2: 1100–1106.

Melling A (1997) Tracer particles and seeding for particle image velocimetry. *Meas. Sci. Technol.* 8: 1406–1416.

Meyers JF, Lee JW, Schawartz RJ (2001) Characterization of measurement error sources in Doppler global velocimetry. *Meas. Sci. Technol.* 12: 357–368.

Morrison GL, Gaharan Jr CA (2001) Uncertainty estimates in DGV systems due to pixel location and velocity gradients. *Meas. Sci. Technol.* 12: 369–377.

Neeteson NJ, Bhattacharya S, Rival DE, Michaelis, D, Schanz D, Schroeder A (2016) Pressure-field extraction from Lagrangian flow measurements: first experiences with 4D-PTV data. *Exp in Fluids* 57: 102.

Nogueira J, Lecuona A, Rodriguez PA (1999) Local field correction PIV: on the increase of accuracy of digital PIV systems, *Exp. in Fluids* 27: 107–116.

Pope SB (2000) *Turbulent Flows*, Cambridge University Press, Cambridge.

Qiu HH, Sommerfeld M (1992) A reliable method for determining the measurement volume size and particle mass fluxes using phase-Doppler anemometry, *Exp. in Fluids*. 13: 393–404.

Qiu HH, Sommerfeld M, Durst F (1994) Two novel Doppler signal detection methods for laser Doppler and phase Doppler anemometry. *Meas. Sci. Technol.* 5: 769–778.

Raffel M, Willert CE, Scarano F, Kähler C., Wereley ST, Kompenhans J (2018) *Particle Image Velocimetry: A Practical Guide*, Springer, Heidelberg.

Ragni D, Ashok A, van Oudheusden BW, Scarano F (2009) Surface pressure and aerodynamic loads determination of a transonic airfoil based on particle image velocimetry. *Meas. Sci. Technol.* 20: 074005

Rife DC, Boortstyn RR (1974) Single tone parameter estimation from discrete time observations. *IEEE Trans. Inform. Theory* 20: 591–596.

Roberts JB, Downie J, Gaster M (1980) Spectral analysis of signals from a laser Doppler anemometer operating in the burst mode, *J Phys. E: Sci. Instrum.* 13: 977–981.

Roehle I (1999) Laser Doppler Velocimetry auf der Basis Frequenzselektiver Absorption: Aufbau und Einsatz eines Doppler Global Velocimeters, DLR research report (Dissertation), ISSN 1434–8454.

Roehle I, Schodl R, Voigt P, Willert C (2000) Recent developments and applications of quantitative laser light sheet measuring techniques in turbomachinery components. *Meas. Sci. Techn.* 11: 1023–1035.

Roisman I, Tropea C (2002) Flux measurements in sprays using PDA techniques, *Atom. & Sprays* 11: 667–699.

Saffman M, Buchhave H, Tanger H (1984) Simultaneous measurement of size, concentration and velocity of spherical particles by a laser Doppler method, *Proc. 2nd Int. Symp. of Laser Anemom. to Fluid Mech.*, Lisbon, Portugal: paper 8.1.

Schmidt J, Völkel R, Stork W, Sheridan JT, Schwider J, Streibl N, Durst F (1992) Diffractive beam splitter for laser Doppler velocimetry, *Opt. Lett.* 17: 1240–1242.

Schröder A, Schanz D, Michaelis, D, Cierpka C, Scharnowski S, Kähler CJ (2015) Advances of PIV and 4D-PTV "Shake-The-Box" for turbulent flow analysis – the flow over periodic hills. *Flow Turb. Combust.* 95: 193–209.

Smeets G, George A (1981) Michelson spectrometer for instantaneous Doppler velocity measurements. *J Phys. E: Sci. Instrum.* 14: 838–845.

Stieglmeier M, Tropea C (1992) Mobile fiber-optic laser Doppler anemometer, *Appl. Opt.* 31: 4096–4105.

Stieglmeier M, Tropea C, Weiser N, Nitsche W (1989) Experimental investigation of the flow through axisymmetric expansions, *ASME J Fluids Eng.* 111: 464–471.

Tennekes H, Lumley JL (1972) *A First Course in Turbulence*. MIT Press, Cambridge, MA.

Unal MF, Lin JC, Rockwell D (1998) Force prediction by PIV imaging: a momentum-based approach. *J. Fluids Struct.* **11**:965–971.

van de Meerendonk RM, Percin B, van Oudheusden BW (2016) Three-dimensional flow and load characteristics of flexible revolving wings at low Reynolds number. In: *18th Int. Symp. Appl. Laser Techn. Fluid Mech.*, 4–7 July, Lisbon, Portugal.

van Oudheusden BW (2013) PIV-based pressure measurement. *Meas. Sci. Techn*. 24: 032001

vom Stein HD, Pfeifer HJ (1969) A Doppler difference method for velocity measurements. *Metrologica* 5 (2): 59–61.

Westerweel J, Nieuwstadt FG (1991) Performance tests on 3-dimensional velocity measurements with a two-camera digital particle-image-velocimeter, *Laser Anemom*. 1: 349–355.

Willert C (1997) Stereoscopic digital particle image velocimetry for application in wind tunnel flows. *Meas. Sci. Technol*. 8: 1465–1479.

Wriedt T, Bauckhage KA, Schöne A (1989) Application of Fourier analysis to phase-Doppler-signals generated by rough metal particles. *IEEE Trans. Instrum. and Meas*. 38: 984–990.

Yeh Y, Cummins, HZ (1964) Localized fluid flow measurements with an He-Ne laser spectrometer. *Appl. Phys. Lett*. 4: 176–178.

Zhang Z, Eisele K (1995) Off-axis alignment of an LDA-probe and the effect of astigmatism on measurements. *Exp. in Fluids* 19: 89–94.

Zhang Z, Eisele K (1996) The effect of astigmatism due to beam refractions on the formation of the measurement volume in LDA measurements, *Exp. in Fluids* 20: 466–471.

FURTHER READING

Adrian RJ (1991) Particle-imaging techniques for experimental fluid mechanics. *Ann. Rev. Fluid Mech*. 23: 261–304.

Adrian RJ, Westerweel J (2011) *Particle Image Velocimetry*, Cambridge University Press, Cambridge, UK.

Albrecht H-E, Borys M, Damaschke N, Tropea C (2003) *Laser Doppler and Phase Doppler Techniques*, Springer-Verlag, Heidelberg.

Benedict LH, Nobach H, Tropea C (2000) Estimation of turbulent velocity spectra from laser Doppler data. *Meas. Sci. Technol*. 11: 1089–1104.

Buchhave P, George WKJr, Lumley JL (1979) The measurement of turbulence with the laser-Doppler anemometer. *Ann. Rev. Fluid Mech*. 11: 443–504.

Czarske J (2006) Laser Doppler velocimetry using powerful solid-state light sources. *Meas. Sci. Technol*. 17:R71–R91.

Meyers JF (1995) Development of Doppler global velocimetry as a flow diagnostics tool. *Meas. Sci. Technol*. 6: 769–783.

Miles RB, Lempert WR (1997) Quantitative flow visualization in unseeded flows. *Ann. Rev. Fluid Mech*. 29: 285–326.

Miles RB, Lempert WR, Forkey JN (2001) Laser Rayleigh scattering. *Meas. Sci. Technol*. 12: R33–R51.

M, Willert C, Kompenhans J (2013) *Particle Image Velocimetry: A Practical Guide*, Springer-Verlag, Heidelberg.

Scarano F (2002) Interative image deformation methods in PIV. *Meas. Sci. Technol*. 13: R1–R19.

Tropea C (1995) Laser Doppler anemometry: recent developments and future challenges. *Meas. Sci. Techn*. 6: 605–619.

Tropea C, Yarin, AL, Foss J (2007) *Springer Handbook of Experimental Fluid Mechanics*, Springer Verlag, Heidelberg.

Tropea C (2011) Optical particle characterization in flows. *Annu. Rev. Fluid Mech*. 43: 399–426.

Van de Hulst HC (1981) *Light Scattering by Small Particles*, Dover Publications, New York.

Westerweel J, Elsinga GE, Adrian RJ (2013) Particle Image Velocimetry for complex and turbulent flows. *Annu. Rev. Fluid Mech*. 45: 409–436.

Willert CE, Gharib M (1991) Digital particle image velocimetry. *Exp. in Fluids* 10: 181–193.

4

Laser Vibrometers

Neil A. Halliwell

CONTENTS

4.1 Introduction 43
4.2 Basic Principles of Laser Vibrometry 44
4.3 Solid-Surface Vibration Measurements 45
    4.3.1 Frequency-Shifting and/or Direction Ambiguity Removal 45
    4.3.2 Optical Geometry/Interferometer 47
    4.3.3 Doppler Signal Processing 48
    4.3.4 Solid-Surface Scattering of Laser Light: Laser Speckle 49
4.4 Limitations of Use 49
    4.4.1 Laser Speckle Effects 49
    4.4.2 Measurements on Rotating Targets 51
    4.4.3 Scanning Laser Vibrometers, Impact Measurements and Practical Considerations 52
4.5 Measurement of Angular Vibration Velocity 54
    4.5.1 Introduction 54
    4.5.2 The Laser Torsional or Rotational Vibrometer 54
    4.5.3 Measurements within a Rotating System 56
    4.5.4 Practical Considerations and Examples of Use 56
References 58

4.1 Introduction

The measurement of vibration of a solid surface is usually achieved with an accelerometer or some other form of surface-contacting sensor. There are, however, many cases of engineering interest where this approach is either impossible or impractical such as measurements on hot, light and rotating surfaces. Practical examples of these are engine exhausts, loudspeaker cones and rotating shafts, respectively. Figure 4.1 shows the vibration velocity contours on the surface of a diamond-impregnated circular saw blade whilst cutting granite [1] and is an excellent example of a difficult measurement problem solved by laser vibrometry. Since the advent of the laser in the early 1960s, optical metrology has provided a means of obtaining remote measurements of vibration which hitherto would have been unobtainable. The methods are based on laser velocimetry.

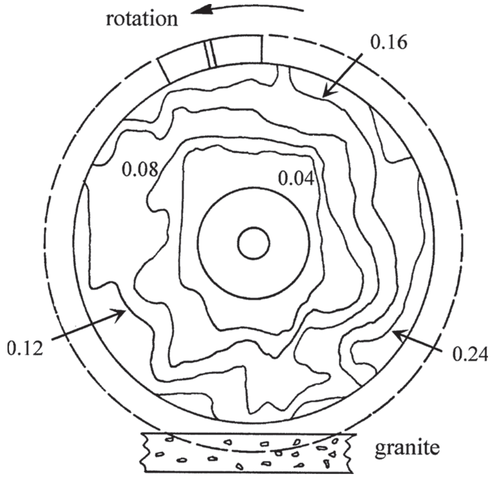


FIGURE 4.1 Stone cutting saw: vibration velocity contours m s^{-1} (rms).

Laser velocimetry is now a well-established technique for non-intrusive time-resolved velocity measurements in fluid flows. The technique relies on detecting the velocity of seeding particles or tracers as they pass through a small volume in the flow. The tracer particles scatter light from the laser beams, and the Doppler frequency shift in the light is detected to provide a measure of the tracer particle velocity. In this way, the flow is essentially sampled by the presence of tracer particles which are present in the volume and are specially chosen both to scatter sufficient light to a photodetector and to follow the flow velocity fluctuations of interest. An obvious practical problem to be considered is the intermittent presence of tracer particles since a Doppler signal processing system which aims to provide time-resolved information must be able to accommodate periods of signal ‘drop-out’ where either no tracers are present in the region or the vector addition of the scattered light from each particle produces a very low amplitude signal from the photodetector. The choice of tracer particles and their spatial distribution in the flow is a major issue in the accurate use of laser velocimetry, and much of the research work published in establishing the technique has concerned the validity of statistical conclusions drawn from a particle measurement.

When laser velocimetry is used to take a velocity measurement of a solid body, the instrument or measurement system is usually referred to as a laser vibrometer. In this situation, the tracer particles can be considered to be 'replaced' by scattering elements within a small surface area of the target body which is illuminated by the incident laser beam. The first major international conference on laser velocimetry was held in Copenhagen, Denmark, in 1975, and when commercial instrumentation, designed for use in fluid flows, appeared in the marketplace, its use for solid-surface velocity measurement was considered to be straightforward in comparison since the intermittency of the Doppler signal was assumed to be no longer a problem, and in general, with appropriate collection optics, it was always possible (even from matt black surfaces!) to collect sufficient light to take a reliable measurement.

In fact, when laser light is scattered from a solid surface, which is rough on the scale of the optical wavelength, a speckle pattern [2] is formed in space in front of the target. The photodetector may sample one or more speckles and this fact brings its own problems when laser vibrometry is used in practice. Rather than the intermittency of tracer particles presenting problems in the fluid flow case, the permanent presence of scattering elements in the solid body case dictates speckle pattern dynamics which, as we shall see, can affect the ultimate performance of the instrument. It is this physical fact which primarily distinguishes the operation of laser vibrometry from its fluid flow counterpart.

In what follows, the basic physical principles of a laser vibrometer measurement are introduced before particular problems relating to solid-body velocimetry measurements are discussed.

4.2 Basic Principles of Laser Vibrometry

The basic working principle of laser vibrometry is the detection of the Doppler shift in coherent light which is scattered from the moving surface and is best explained as an extension to the general fluid flow case shown in Figure 4.2. With reference to this figure, when a tracer particle with a velocity vector $\boldsymbol{U}$ scatters light of wavelength λ in a direction $\boldsymbol{K}_2$ from an incident laser beam in a direction $\boldsymbol{K}_1$, where both $\boldsymbol{K}_2$ and $\boldsymbol{K}_1$ are unit vectors, the scattered light undergoes a Doppler shift in frequency f_D given by [3]

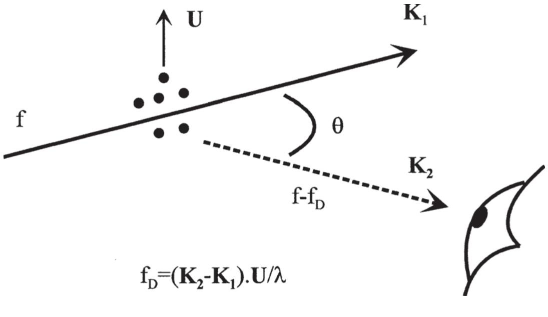


FIGURE 4.2 Doppler shift in scattered light (fluid flow).

$$f_D = (\boldsymbol{K}_2 - \boldsymbol{K}_1) \cdot \boldsymbol{U}/\lambda. \tag{4.1}$$

Laser vibrometry involves directing an incident laser beam at the target surface of interest, and an instrument will collect light which is directly backscattered off the target surface as shown in Figure 4.3. With reference to this figure, the Doppler shift f_D is given by

$$f_D = |\, 2\boldsymbol{K}_1 \cdot \boldsymbol{U} \,/\, \lambda = \frac{2|\boldsymbol{U}|\cos\theta}{\lambda} \tag{4.2}$$

since now $\boldsymbol{K}_1 = -\boldsymbol{K}_2$.

In this way, by demodulating and tracking the changing value of f_D, it is possible to produce a time-resolved measurement of the component of solid surface velocity which is in the direction of the incident laser beam.

The scattered light beam from the target surface has a frequency of typically 10^{15} Hz which cannot be demodulated directly. The Doppler frequency f_D is detected using an optical interferometer by mixing the scattered beam coherently with a reference beam derived from the same laser source onto the surface of a photodetector. The photodetector responds to a time average of the intensity of the total light collected, and this form of non-linear detection produces a heterodyne or 'beat' in the current output whose frequency is equal to the difference in frequency between the two beams [3]. This arrangement is shown schematically in Figure 4.4 in Michelson interferometer geometry. With reference to this figure, the incident laser beam

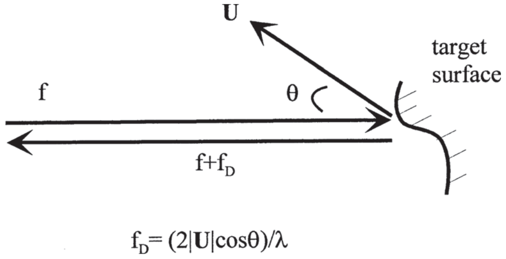


FIGURE 4.3 Backscatter geometry for laser vibrometry.

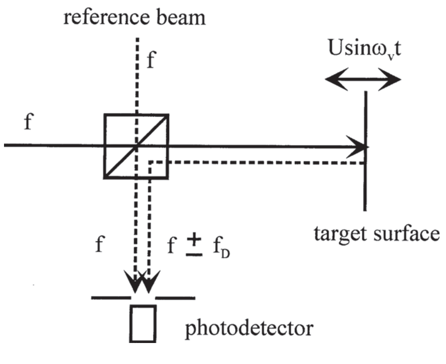


FIGURE 4.4 Reference beam heterodyne (Michelson interferometer).

has a frequency *f*, and without the loss of generality, the target surface is shown vibrating with a sinusoidal velocity $U \sin \omega_v t$ in a direction parallel to the incident laser beam where ω_v is the angular vibration frequency.

The light intensity incident *at a point* on the photodetector surface $I(t)$, neglecting polarization effects, is given by [3]

$$I(t) = I_R + I_T + 2\sqrt{I_R I_T}\cos[2\pi f_D t + \phi_R - \phi_T] \tag{4.3}$$

where I_R is the reference beam intensity, I_T is the target beam intensity, ϕ_R is the constant reference beam phase and ϕ_T is the target beam phase, $f_D = (2U \sin \omega_v t)/\lambda$ is the 'instantaneous' Doppler frequency shift in Hz (strictly the phase change due to the Doppler effect at time t is $\int f_D \mathrm{d}t$).

This arrangement is ambiguous in the measurement of the direction of the vibration velocity. When the Doppler frequency shift is non-zero, it is not possible to distinguish whether the target is moving towards or away from the detector. Figure 4.5 shows the spectrum of the photodetector output. Since the vibrometer provides a time-resolved measure of surface velocity by tracking the Doppler frequency, it is also clear that the signal will be discontinuous with this arrangement and, thus, not generally possible to decode properly. Vibration engineers require a measure of amplitude *and* phase of the target velocity, and this is provided by frequency pre-shifting the reference beam.

Figure 4.6 shows Michelson interferometer geometry where the reference beam has been frequency-shifted by a constant amount f_R. In this case, the photodetector output is given by

$$I(t) = I_R + I_T + 2\sqrt{I_R I_T}\cos[2\pi(f_R - f_D)t + \phi_R - \phi_T]. \tag{4.4}$$

With reference to the figure, the output spectrum of the photodetector now shows how the frequency shift provides a carrier frequency in the photodetector output which the target surface velocity frequency modulates. In this way, frequency demodulating and tracking the photodetector output signal provides a time-resolved measurement of target surface velocity.

Laser vibrometer systems assembled for laboratory use from bulk components or purchased as a commercial instrument all work on the physical principles so far described. They will differ in the choice of optical geometry, means of frequency-shifting and/or direction ambiguity removal and electronic design of Doppler signal processor. In the next section, practical considerations governing this choice are considered.

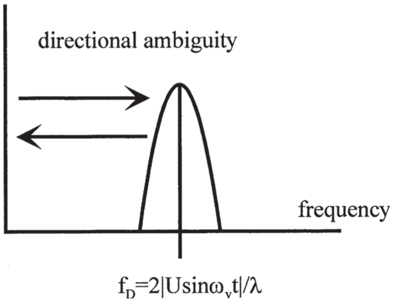


FIGURE 4.5 Photodetector output spectrum.

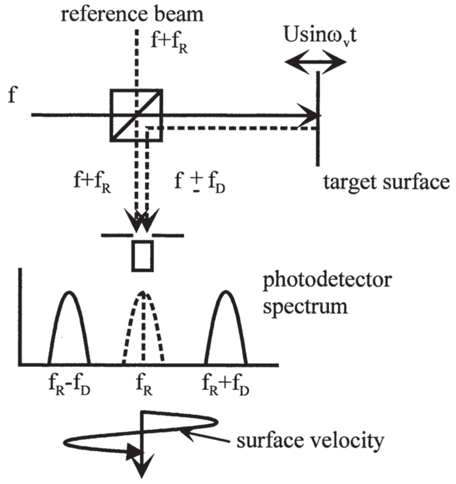


FIGURE 4.6 Frequency-shifting (Michelson interferometer).

4.3 Solid-Surface Vibration Measurements

4.3.1 Frequency-Shifting and/or Direction Ambiguity Removal

Various means of frequency-shifting have been employed in laser vibrometer systems as the technology has developed and these include rotating scattering discs [4], Kerr cells [5], rotating diffraction gratings [6], piezoelectric elements [7], fibre optic modulators [8], Bragg cells [9] and laser diode current modulation [10].

An early, simple and robust frequency shifter in a vibrometer was provided by a rotating scattering disc [11,12] and this is shown in Figure 4.7. With reference to this figure, the normal to the plane of the disc is at a small angle θ to the incident laser beam so that scattering elements passing through the beam have a velocity component in the direction of the latter. The magnitude of the frequency shift is readily controlled by the speed of rotation and/or the radial position of the incident beam. In this way, the reference beam takes the form of a scattered light beam in a solid angle which is dictated by the limiting aperture (usually at the photodetector) in the light collection system. Shifts of typically 1 MHz and higher are easily provided. The disc is covered with retro-reflective tape to ensure adequate intensity in direct backscatter. Frequency-shifting in this way provides a self-aligning property, useful in large-scale bulk component systems, which is not integral to other forms of frequency-shifting. This is at the expense of a noise floor performance which, when measured on a stationary target, displays speckle noise peaks in the form of a periodogram with a fundamental at the rotation frequency.

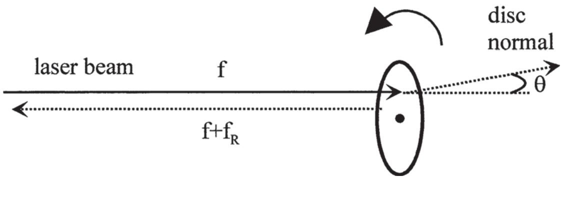


FIGURE 4.7 Frequency-shifting by scattering disc.

The minimum vibrational level measurable at these frequencies is reduced. Commercial instrumentation [13] which utilized this form of shifting provided a choice of disc speeds to avoid this problem.

Another means of frequency-shifting is provided by a rotating diffraction grating [6]. This is shown in a Michelson interferometer in Figure 4.8. With reference to this figure, a small glass disc (; 30 mm in diameter) is radially etched so that the incident laser beam is subject to periodic thickness variations as the disc is rotated. In this way, periodic changes in optical path length on a spatial scale comparable with the laser wavelength provide a moving phase diffraction grating. The light is diffracted into distinct orders which are frequency-shifted in proportion to the disc speed which is easily controlled. Early commercial versions of this frequency shifter [14] provided shifts of *circa* 1 MHz. Choice of optical geometry in the vibrometer dictates which orders are used. For a normal-to-surface vibration measurement, it is usually the first order which is chosen. The etched profile is designed to ensure adequate intensity in this order.

The noise floor of this vibrometer when measured on a stationary target surface is a periodogram with a fundamental frequency at the disc rotational speed. This is due to the machining process in manufacturing the disc. Consequently, as with the scattering disc, the minimum measurable vibration level at these frequencies is reduced. The rotating grating provides a flexible tool in bulk component vibrometer systems. They are inherently fragile and experienced users treat them with care!

To date, the most commonly used form of frequency-shifting in commercial vibrometers is the Bragg cell. This was employed in the first commercially available vibrometer in 1975 [9]. A Bragg cell design is shown in Figure 4.9. With reference to this figure, the incident laser beam traverses a transparent medium (usually a crystal) in which an ultrasonic wave is travelling normal to the beam direction. Density variations, which are on the same spatial scale as the optical wavelength, produce refractive index variations and, in this way, provide a moving phase diffraction grating. The incident beam is diffracted into distinct orders each undergoing a frequency shift which is a direct multiple of the frequency of the ultrasonic wave (typically 40 MHz). Normally, the cell is designed so that the first-order diffracted beam is used as the reference beam which emerges at an angle to the incident beam and may, therefore, require alignment compensation in some geometries. The solid-state nature of the Bragg cells makes them the first choice as frequency shifters in commercial instruments but their presence does add to the expense since they require additional electronic and optical components.

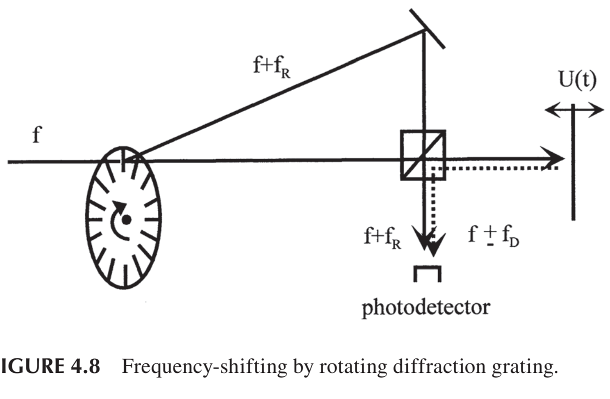


FIGURE 4.8 Frequency-shifting by rotating diffraction grating.

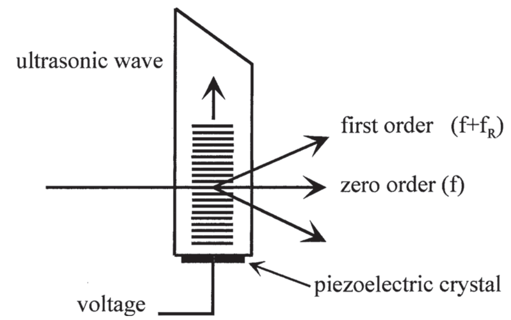


FIGURE 4.9 The Bragg cell.

As an alternative to direct frequency-shifting, directional ambiguity in the Doppler signal can be removed by quadrature detection [15]. This is so called because polarization optics are used to provide two separate detection channels in which the reference beam in one channel is π/2 out-of-phase with the other. In this way, the sine-cosine relationship between the detector outputs provides for a deterministic response in their relative phase when the target surface changes direction. This technique is normally used in a Michelson interferometer geometry and one such method suggested by Reibold and Molkenstruck [16] is shown is Figure 4.10. With reference to this figure, the reference beam passes twice through a $\lambda/8$ retardation plate after reflection from a reference mirror to return as a circularly polarized beam before recombination with the signal beam at the beamsplitter. The combined beams are split by a Wollaston prism (polarizing beamsplitter) whose axis is at 45° to the original polarization direction of the laser. In this way, two signal channels are created where the reference beams are in quadrature, thus the signals from the photodetectors have the desired sine/cosine relationship. Quadrature detection has been incorporated into some commercially available vibrometer systems [17].

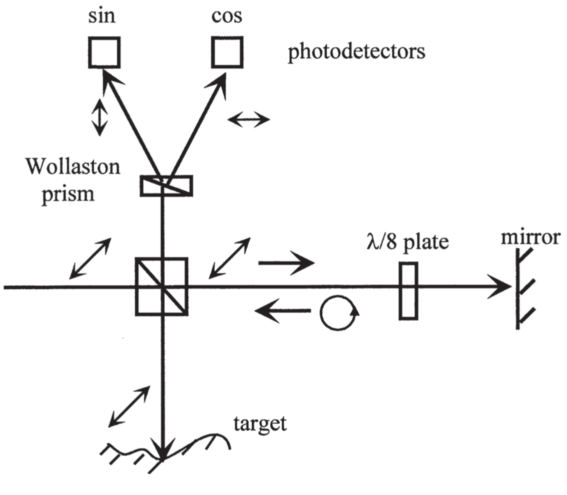


FIGURE 4.10 Quadrature detection. (After Reibold and Molkenstruck [16].)

4.3.2 Optical Geometry/Interferometer

The two most popular choices of optical interferometers are the Michelson and Mach–Zender. The Michelson is shown in Figures 4.7, 4.8 and 4.10. A Mach–Zender interferometer, which utilizes a Bragg cell as a frequency-shifting element, is shown in Figure 4.11. With reference to this figure, a polarizing beamsplitter (PBS1) splits the incident beam into target and reference beams. The target beam is transmitted by a second polarizing beamsplitter (PBS2). Assuming no change in polarization after scattering from the target surface, two passes through the $\lambda/4$ retardation plate rotates the plane of polarization through 90° so that the target beam is then reflected by PBS2. This offers the additional advantage that the laser source is isolated from the target light which avoids feedback problems with the laser. A final beamsplitter combines the target and reference beams after the latter has been frequency-shifted by the Bragg cell. The beamsplitter produces two separate channels for detection which are π out-of-phase. This is an advantage since taking the difference of the two detector outputs means the detection system is insensitive to stray coherent light and to intensity modulation of the reference beam. The latter is important when relatively high reference beam-to-scattered light intensity ratios are used to obtain increased sensitivity when the scattered light intensity is low. Modulation of the reference beam intensity can occur through power supply ripple to the laser and inter-mode-beating effects [3]. The level of this is typically 1% for a He-Ne laser.

An optical geometry which provides for direct measurement of the relative vibration between two surfaces was suggested by Selbach and Lewin [18]. It is a differential fibre optic vibrometer based on a Mach–Zender interferometer and is shown in Figure 4.12. With reference to this figure and in comparison with Figure 4.11, the initial beamsplitter (BS1) is now no longer polarized and the mirror in the reference beam path is replaced with a polarizing beamsplitter (PBS3) such that the reference beam is transmitted and launched into a polarization-conserving fibre. In this way, the reference beam provides for a second 'target' beam and light scattered from the reference surface returns to be reflected by PBS3 after two passes through a $\lambda/4$ retardation plate which rotates the plane of polarization through 90°. The reference beam is then frequency-shifted by the Bragg cell. The original target beam is also coupled to a polarization-preserving fibre for transmission to the target surface. Detection of the signal and

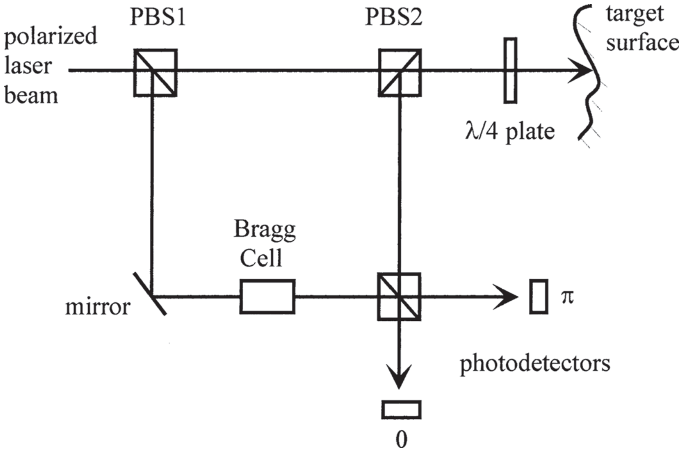


FIGURE 4.11 Mach–Zender interferometry design.

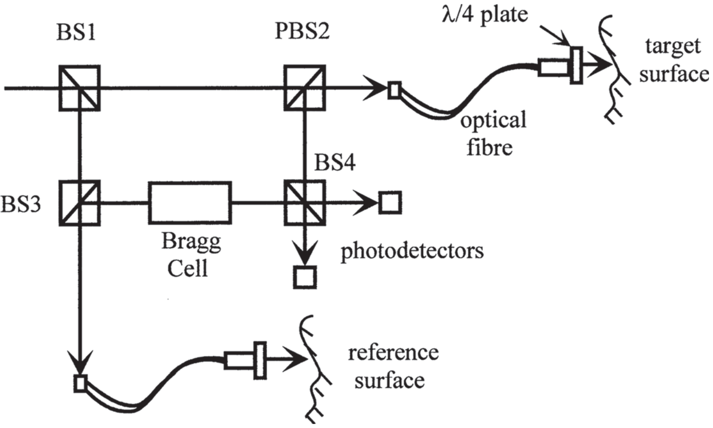


FIGURE 4.12 A differential fibre vibrometer. (After Selbach and Lewin [18].)

reference light is as described for the classical Mach–Zender interferometer.

This vibrometer measures the difference in vibration level between the target and reference surfaces and has been developed commercially [19]. If the reference surface is replaced by a mirror, it can function as a single-point vibrometer as per Figure 4.11. Use of fibre optics provides an advantage of easy access to remote surfaces and specially designed optical heads provide higher numerical aperture light gathering to ensure adequate intensity in the scattered target and reference beams. Care should be taken to avoid extraneous phase and/or amplitude disturbances produced in the light transmitted by the fibres in either arm. These can be caused through local refractive index variations due to noise, vibration, heat, etc., or through disturbance by electromagnetic fields. These disturbances will affect the signal-to-noise ratio and make the vibrometer microphonic. An all-fibre design which avoids these problems was suggested by Laming et al. [20].

An optical geometry which provides for in-plane vibration measurement (perpendicular to the optical axis) is that of the cross-beam laser velocimeter or differential Doppler technique [21]. The target surface is arranged to be in the intersection region of the two laser beams as shown in Figure 4.13. With reference to this figure, the 'beat' frequency f_{D} in the photodetector output is given by

$$f_{\mathrm{D}} = \frac{2U}{\lambda}\sin\frac{\theta}{2} \tag{4.5}$$

where U is the surface velocity normal to the axis of symmetry of the laser beams, λ is the wavelength of the laser light and θ is the included angle between the incident laser beams.

Typically, the intersection volume of the laser beams is ellipsoidal and 1 mm in length along the major axis. Care must be taken to ensure that the surface remains within the volume during a measurement to avoid loss of signal. The cross-beam velocimeter has been used to measure the tangential surface velocity and, hence, rotational speed and/or torsional oscillations of rotating circular cross-section shafts [22]. Solid body oscillations of the shaft can preclude this measurement, however, and use of the cross-beam velocimeter for torsional vibration measurement has been superseded by the invention of the laser torsional vibrometer.

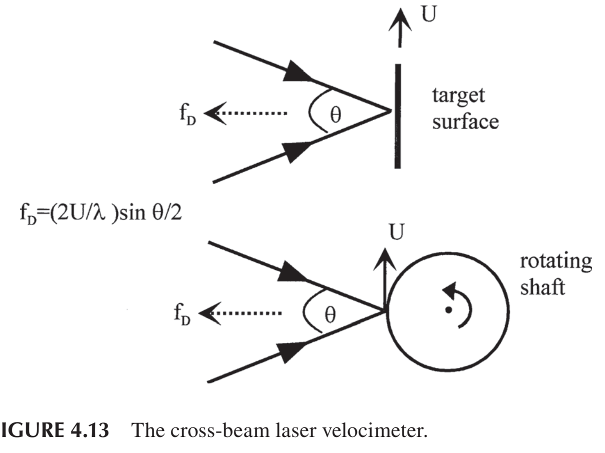


FIGURE 4.13 The cross-beam laser velocimeter.

4.3.3 Doppler Signal Processing

The choice of processing scheme is, in general, dictated by the characteristics of the Doppler signal itself, and in laser vibrometry applications, Doppler signals are continuous but can be subject to periods of very low amplitude due to speckle effects. These periods of low amplitude are analogous to 'drop-outs' in the fluid flow application of laser Doppler velocimetry.

The continuous nature of the Doppler signal in laser vibrometry lends itself to frequency tracking and a typical arrangement is that suggested by Wilmshurst [23] and is shown schematically in Figure 4.14. With reference to this figure, a voltage-controlled oscillator (VCO) is used to track the Doppler signal and is controlled via a feedback loop. A mixer at the input stage produces an 'error' signal between the Doppler and VCO frequencies which is bandpass filtered and weighted before being integrated and used to control the oscillator to drive the error to a minimum. The feedback loop has an associated 'slew rate' which limits the frequency response of the processor. With respect to the Doppler signal, the tracker acts effectively as a low-pass filter which outputs the VCO voltage as a time-resolved voltage analogue of the changing frequency and, hence, vibration velocity. Some trackers incorporate sophisticated weighting networks which tailor the control of the VCO according to the signal-to-noise ratio of the Doppler signal. The simplest form of this network will hold the last value of the Doppler frequency being tracked should the amplitude of the signal fall below a pre-set value. In this way, the real Doppler signal effectively 'drops out' and consideration must be given to the statistics of what is, therefore, an essentially sampled output especially when high-frequency vibration information of the order of the 'drop-out' period (typically several microseconds) is required.

Electronic designs of Doppler signal processors vary and each will compensate in some way for loss of Doppler signal amplitude in order to produce a time-resolved output. Commercially available laser vibrometer systems will only specify the performance of the processor in the absence of speckle effects where the Doppler signal amplitude is maintained at a high level. This is commensurate, in practice, with the target surface moving strictly parallel to the incident laser beam. Such 'ideal' signals can be demodulated with high resolution, accuracy and frequency response. Commercially available laser vibrometers specify low-vibration velocity limits of

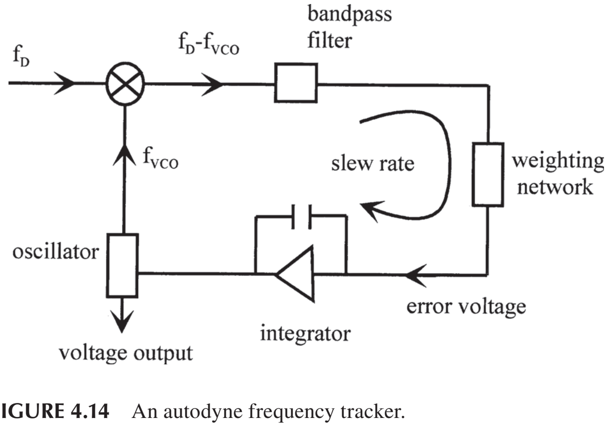


FIGURE 4.14 An autodyne frequency tracker.

typically fractions of a micrometre per second (bandwidth-dependent) and frequency responses of MHz.

It should be remembered that the Doppler frequency is proportional to surface velocity and, as such, may relate, in practice, to high surface displacement amplitudes at low vibration frequencies or *vice versa* for the same value of Doppler frequency. At very low vibration frequencies, processors may operate in 'zero-crossing' or 'fringe-counting' mode [21]. They then produce a time-resolved voltage output proportional to surface displacement rather than velocity. Commercial instruments specify displacement resolution limits of nanometres.

4.3.4 Solid-Surface Scattering of Laser Light: Laser Speckle

For a laser vibrometer to operate, it is necessary to ensure sufficient intensity of light is backscattered along the path of the incident laser beam. With reference to Equation (4.4), the target intensity I_T must be large enough to ensure the amplitude of the ac term provides an adequate signal-to-noise ratio for the Doppler signal processor. For this reason, highly polished and reflective targets cause problems since they obey the law of reflection (angle of incidence = angle of reflection) with little diffuse scattering occurring. In laser vibrometry applications, the target surface should scatter the incident light diffusely.

An ideal diffuse scattering surface obeys Lambert's cosine law of diffusion given by

$$W(\theta) = \frac{\Gamma}{\pi} W_i \cos\theta \tag{4.6}$$

where $W(\theta)$ is the scattered light power per unit solid angle in a direction θ relative to the normal to the surface, W_i is the incident light power and the constant Γ varies from unity for a white surface and less for absorbing surfaces which are typically grey or black. Most target surfaces in engineering behave somewhere between the two extremes of specular reflection and ideal diffuse scattering. A large advantage can be obtained if the surface can be treated with retroflective tape or paint. In this case, light is always scattered back along the path of the incident beam almost independently of the surface orientation. Such surface treatment contains small spheres of high refractive index embedded in a medium, and these return the light to the source over an angle of a few degrees. Reflections from sphere to sphere within the medium still cause a laser speckle effect which is introduced as follows.

When a laser beam is incident on a target whose surface is rough on the scale of the optical wavelength, the irradiated area behaves as a dense collection of individual scattering elements. Each element scatters the coherent light so that, at a point in space in front of the target, the contributions from each element add together constructively or destructively. In this way, if the intensity distribution is observed on a screen in front of the target, a laser speckle pattern [2] is formed. The pattern consists of a random array of bright and dark areas (speckles) and a typical example is shown in Figure 4.15. With reference to this figure, each speckle represents a small area of correlation where it can reasonably be assumed that the resultant amplitude and phase from the vector sum of the scattered light from each element is constant. In practice, speckles form approximately ellipsoidal volumes in space and the picture shown in the figure depicts a cross-sectional intensity map from each adjacent speckle volume.

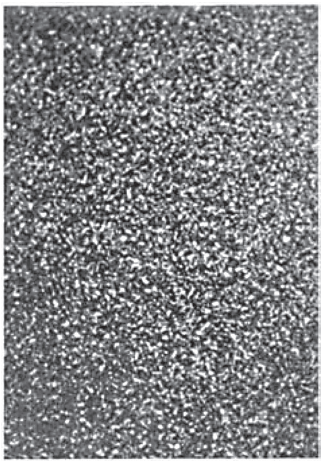


FIGURE 4.15 A laser speckle pattern.

The photodetector in a laser vibrometer samples the speckle pattern generated by the target surface. Speckle size is inversely proportional to the dimensions of the illuminated area and dictated by the limiting aperture in the light collection and/or imaging system used [2]. Most commercially available laser vibrometers are designed to focus the laser beam onto the target and operate by sampling a single bright speckle in the pattern. With reference to Equation (4.4), this means that, in practice, the target intensity term I_T is maximized so that the signal-to-noise ratio is high. This situation is an optimum for laser vibrometer operation and is commensurate with instrument specifications quoted in commercial literature. The effect of the initial speckle, which is sampled by the detector, changing during a measurement is discussed in the next section.

4.4 Limitations of Use

4.4.1 Laser Speckle Effects

It is the fact that the photodetector in a laser vibrometer samples a laser speckle pattern which distinguishes laser vibrometer operation from that of laser Doppler anemometry in the fluid flow case. One or more speckles may be sampled by the detector depending upon the optical geometry used and/or the focusing/collection optics used in a particular system or instrument. Most commercial instruments operate by sampling just one bright speckle which is heterodyned with a uniform frequency-shifted reference wave. Initially, the instrument is aligned with a stationary target so that a bright speckle is sampled by the photodetector. Since a speckle is a correlation region in space in which it can be reasonably assumed that the amplitude and phase of the light remains constant if, during a measurement, the sampled speckle does not change then the vibrometer operates as a classical interferometer where the *rate of change* of the speckle phase modulates the carrier frequency provided by mixing with the frequency-shifted reference beam.

In this situation, we re-write Equation (4.4) describing the intensity $I(t)$ at a point on the detector as

$$I(t) = I_{\mathrm{R}} + I_p + 2\sqrt{I_{\mathrm{R}} I_p}\cos[2\pi(f_{\mathrm{R}} - f_{\mathrm{D}})t + \phi_{\mathrm{R}} - \phi_p] \tag{4.7}$$

where the suffix 'p' now identifies that it is the pth speckle which is contributing to the detector output at this point on its surface. I_p is the intensity of the speckle and ϕ_p is the speckle phase.

In practice, the photodetector responds to an instantaneous sum of $I(t)$ over its active area A and we must, therefore, consider the spatial distributions of I_{R} and I_p and their respective phases. For convenience, we can assume that the reference beam intensity and phase are uniform in space and we need only to consider the speckle pattern distribution. The detector output current $i(t)$ is, therefore, given by

$$\begin{aligned} i(t) = \beta I_{\mathrm{R}} A + \beta \int_A I_p(a)\ \mathrm{d}a \\ +2\beta \int_A \sqrt{I_{\mathrm{R}} I_p(a)}\cos[2\pi(f_{\mathrm{R}} - f_{\mathrm{D}})t + \phi_{\mathrm{R}} - \phi_p(a)]\mathrm{d}a \end{aligned} \tag{4.8}$$

where β is a photoelectric constant, $I_p(a)$ is the speckle intensity distribution on the photodetector and $\phi_p(a)$ is the speckle phase distribution on the photodetector.

Examination of the photodetector output shows that when the speckle intensity and phase distributions change with time during a measurement, spurious information will be contributed when $i(t)$ is frequency-demodulated. The major noise source is due to the phase term ϕ_p becoming time-dependent and contributing a frequency modulation $\mathrm{d}\phi_p/\mathrm{d}t$ which is indistinguishable from the change in Doppler frequency associated with the intended measurement. For a commercial instrument which detects just one speckle on the photodetector, it must be ensured that the same speckle remains on the photodetector during a measurement to avoid this problem. The contribution of speckle noise termed 'pseudo vibrations' has been examined in detail Ref. [24], and readers are referred to this paper for a full discussion.

In brief, if the surface moves strictly parallel to the incident laser beam during an intended normal-to-surface vibration measurement, then the speckle pattern characteristics of intensity and phase can be assumed to be constant. This is because the speckles form approximately ellipsoidal volumes in space, and the detector plane effectively samples at different positions along the major axis of each volume during this form of target motion. When correctly aligned on an initially stationary target surface, an instrument operating by collecting just one bright speckle will essentially 'stay with' this speckle during the measurement.

If, however, the target surface tilts, moves in-plane or rotates during a measurement, the speckle characteristics will change. When the scattering elements in the laser spot rotate locally due to surface tilt, the associated speckle pattern on the detector moves in sympathy and this is usually the dominant noise source. If a significant change in population of scattering elements occurs (perhaps due to in-plane target motion), then the speckle pattern will evolve or 'boil' in sympathy to produce noise. In practice, both mechanisms contribute pseudo-vibrations in the laser vibrometer output and which one dominates depends upon the measurement attempted.

It is important to note that commercial instrument manufacturers quote performance for laser vibrometers when operating on a stationary target. In practice, this noise performance will only be obtained when the bright speckle detected during the alignment of the instrument on a stationary target does not change during the measurement. This point is given emphasis by considering the design of a laser vibrometer which utilizes a rotating scattering disc as a frequency-shifter [11]. The noise floor spectrum, for a stationary target and a tripod-mounted instrument, is as anticipated, a periodogram at a 5-Hz fundamental disc frequency generated by speckle noise as shown in Figure 4.16. This is comparable with the noise floor spectrum which would be produced by use of a laser vibrometer with frequency-shifting by a Bragg cell when the vibrometer was taking a measurement from a target which was rotating simply at 5 Hz. This emphasizes the point that noise floors quoted by commercial instrument manufacturers do not apply to

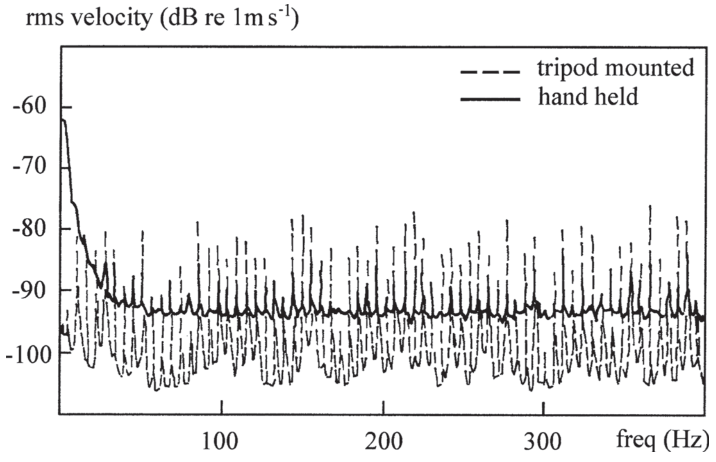


FIGURE 4.16 Vibrometer noise floor (5 Hz scattering disc).

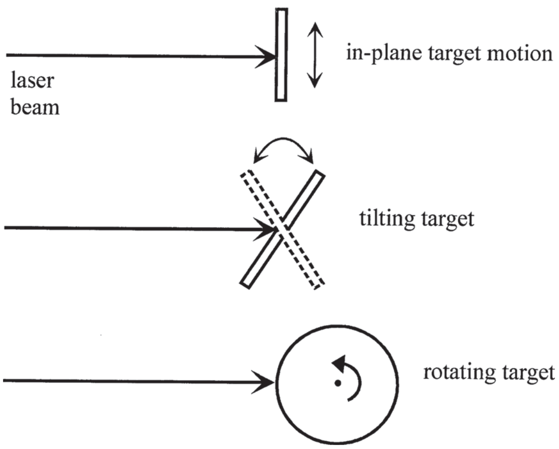


FIGURE 4.17 Target motions for speckle noise generation.

every measurement situation but rather depend on the speckle pattern dynamics in each case.

Commercial instrument specifications of dynamic range, frequency response and accuracy are also measured when the speckle detected during alignment remains on the photodetector during the measurement. These specifications may change significantly if this situation is not maintained and speckle noise is produced. Figure 4.17 identifies practical situations where speckle noise is present and careful consideration is needed in interpreting the vibrometer data. With reference to this figure, it is clear that speckle noise is unavoidable during use of an in-plane laser vibrometer. It is also inherent in the vibrometer output if the target surface tilts or rotates. The noise peak heights in the spectrum of a laser vibrometer output on a rotating target can be *circa* 1 mm s^{-1} rms at the rotational frequency and higher order harmonics [24]. Unfortunately, in practice, this noise is often linked to the vibration measurement frequency of interest, and it is in this situation where a degree of engineering judgement is needed in interpreting the results.

It is interesting to note the effect on speckle noise of hand-holding a laser vibrometer. This is shown in Figure 4.16 where the laser beam from the hand-held instrument [11] is incident on a stationary target. Operator body movement is known to be restricted to below 30 Hz [25], and this is shown clearly in the result. It can be seen that operator movement has the effect of causing the speckle pattern from the target to change so that the effect of the periodic repeat is weakened. This causes a reduction in the height of the speckle noise peaks and a corresponding slight increase in the *average* noise level.

4.4.2 Measurements on Rotating Targets

Laser vibrometer measurements on rotating targets are often quoted as an ideal application due to the inherent non-contact capability they offer. They are not, however, straightforward to interpret in the majority of situations. There are two problems to consider. The first has been discussed in the preceding Section 4.4.1 in that the spectrum of the noise floor of a laser vibrometer, in this measurement situation, takes the form of a periodogram with a fundamental frequency at the rotational speed. The pseudo-vibration speckle noise peaks can be at a level of mm s^{-1} at these frequencies [24]. Clearly if, as an example, radial motion of a rotating shaft is the intended measurement, the noise floor is severely affected at the rotation frequency and higher order harmonics.

The second problem concerns the cross-coupling between the intended solid body vibration velocity measurement and the angular motion. Figure 4.18 shows a solid body of arbitrary cross section rotating with angular velocity $\omega(t)$ about an axis within the body. Without loss of generality, the point O defines an origin on the axis which itself is vibrating with a velocity $\boldsymbol{V}(t)$. The laser beam from the vibrometer whose direction is defined by the unit vector $\hat{s}$ is incident on the body at the point P. The position vector of *P* with respect to the origin is $\boldsymbol{r}_\mathrm{p}(t)$ and can be written as

$$\boldsymbol{r}_\mathrm{p}(t) = \boldsymbol{r}_\mathrm{o}(t) + p(t)\hat{\boldsymbol{s}} \tag{4.9}$$

where $\boldsymbol{r}_\mathrm{o}(t)$ is the distance from O to a *fixed point* A on the laser beam axis and $p(t)$ is the distance of A from P. Changes in $p(t)$ occur due to body rotational, tilt and translational motions. The total velocity $\boldsymbol{V}_\mathrm{p}(t)$ of the scattering elements which are illuminated at P is given by

$$\boldsymbol{V}_\mathrm{p}(t) = \boldsymbol{\omega}(t) \times \boldsymbol{r}_\mathrm{p}(t) + \boldsymbol{V}(t). \tag{4.10}$$

The laser vibrometer (see Equation (4.2)) measures the velocity component $V_\mathrm{l}(t) = \hat{\boldsymbol{s}}\boldsymbol{V}_p(t)$ in the direction of the incident beam. Substituting for $\boldsymbol{r}_\mathrm{p}(t)$ from Equation (4.9) the triple scalar product is identically zero to give

$$V_\mathrm{l}(t) = \hat{\boldsymbol{s}} \cdot (\boldsymbol{\omega}(t) \times \boldsymbol{r}_\mathrm{o}(t)) + \hat{\boldsymbol{s}} \cdot \boldsymbol{V}(t). \tag{4.11}$$

This result shows that the vibrometer measurement is independent of the cross-sectional shape of the target. In addition, the intended solid body vibration measurement $\hat{s}$ $\boldsymbol{V}(t)$ is 'contaminated' by the measurement of the cross-coupling between the rotational motion $\omega(t)$ and the solid body vibration displacement defined by $\boldsymbol{r}_\mathrm{o}(t)$. It is also important to note that changes in rotational speed and/or local tilt of the rotational axis will also affect the intended measurement. The problem is how to distinguish the solid body vibration motion and the use of two laser vibrometers arranged to take orthogonal measurements in the rotating target as shown in Figure 4.19 has provided

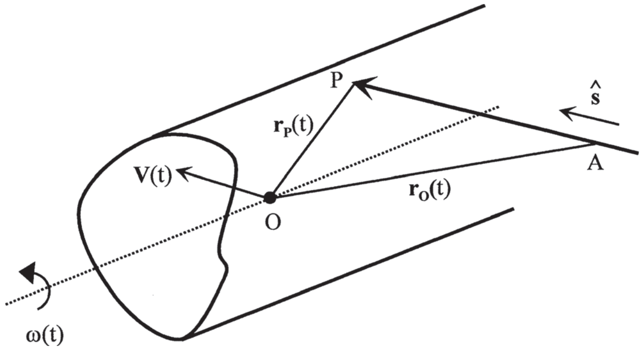


FIGURE 4.18 Vibrometer measurements on rotating targets.

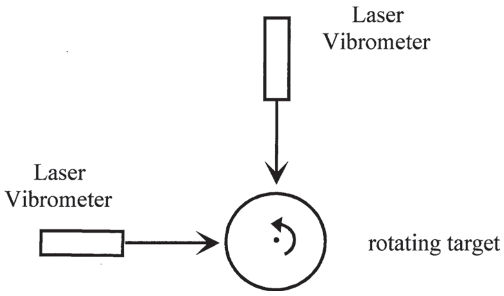


FIGURE 4.19 Use of two orthogonal laser vibrometers.

a solution for certain measurement conditions [26,27]. In particular, where the angular velocity vector $\omega(t)$ can be assumed to be constant, accurate vibration spectra for solid body motion at non-synchronous frequencies can be achieved. A practical solution in this situation is suggested by Bell and Rothberg [28].

The same authors [29] have developed a six degrees of freedom model for laser vibrometer measurements on a rotating shaft. The shaft axis of rotation is allowed to vibrate angularly through small angles during a measurement, as well as undergo torsional and translational vibrations. It is shown that, in the general case, it is not possible to measure radial, axial or bending vibrations of the shaft *directly* by either a single or combination of laser vibrometer measurements. Only unambiguous measurement of torsional vibration is possible using the laser torsional vibrometer described in Section 4.5.2.

4.4.3 Scanning Laser Vibrometers, Impact Measurements and Practical Considerations

Laser vibrometers offer excellent time saving when a spatial array of vibration measurements are needed. An example of this is the need for a survey of an automotive engine in a test cell installation. Use of laser vibrometry avoids the need for the laborious task of drilling and tapping holes for accelerometer mounts and, therefore, significantly reduces test cell usage time, thus offering high cost savings. It is possible to hand-hold a laser vibrometer to save time in the knowledge that vibration spectra obtained will only be significantly affected below 30 Hz (see Section 4.4.1). Hand-holding the instrument will add speckle noise to the output spectrum, and the level in the bandwidth of interest can be ascertained prior to commencing the measurement. If significant, compared with the anticipated measurement level, the instrument should be tripod-mounted and vibration isolated from the target (see later in this section).

An extremely useful extension of laser vibrometry has been provided by the ability to scan the laser beam over the target surface area. This can be achieved, with high positional accuracy, using two moving mirrors driven by galvanometric actuators; vibrometers that include this facility are termed scanning laser vibrometers. Commercially available instruments [30,31] offer scan angles of up to 40° × 40° and working distances of up to 200 m subject to surface treatment. They incorporate advances made in video imaging and automatic computer control. Systems can be mouse-driven in a Windows PC format and are able to superpose colour-coded vibration data on frozen images of the target. Further to this, the scanning laser vibrometer's ability to obtain time-resolved velocity data from a large array of closely positioned test points on the target surface permits calculations of modal parameters from frequency response functions so that animated displays of target vibration can be produced using modal analysis software. This has produced an expanding area of interest in the use of scanning laser vibrometers for modal analysis applications and the refinement of finite element analysis code predictions [32,33]. It is important to note, however, that whilst scanning a target area, the laser beam must be brought to rest momentarily, when data are actually acquired, in order to avoid phase noise due to laser speckle effects described in Section 4.4.1. Stanbridge and Ewins [34,35], however, have successfully demonstrated that important modal analysis information can be obtained by analyzing the modulation in the vibration velocity output, caused by the changing vibration pattern, as the beam is continuously scanned across the target without hesitation.

Laser vibrometry is often a singular solution to problems of velocity measurements of impacting bodies where very high levels of acceleration can occur. It avoids the need for special, high '*g*', shock accelerometers, the use of which, in many cases, is often precluded. Figure 4.20 shows the lateral deformation (normal to line of flight) of a golf ball measured during the brief moment of impact with a golf club [36]. There are two major considerations in attempting measurements of this type. The first is to ensure that the velocity to be measured lies within the dynamic range of the instrumentation used, since the velocity amplitudes measured may well be in excess of those normally encountered in more routine vibration measurements. The second, more subtle, consideration concerns the ability of the Doppler signal processor to follow a sudden change in the velocity of the target, i.e. a sudden acceleration. In impact situations, it can be the 'slew rate' of the processor (identified in Section 4.3.3) which is a limiting factor. With reference to Figure 4.20, the highest acceleration of the golf ball surface corresponds to a velocity change of $7\,\mathrm{m\,s^{-1}}$ in $30\,\mu\mathrm{s}$ giving an acceleration of $233\,\mathrm{km\,s^{-2}}$ which was within the slew rate of the laser vibrometer used. In practice, the notional time over which the largest value of acceleration is recorded can be used to define an upper frequency limit. This estimated frequency should be well within the frequency response of the laser vibrometer for an accurate measurement [37].

In addition, it is important to note that when the vibrometer is not measuring the velocity of the *same* scattering elements during the period of the measurement, an account must be taken of this in attempting to define target deformation. This is the case for the golf ball deformation results presented in Figure 4.20 where the ball translates normal to the incident laser beam during the impact measurement. This effect was compensated for by taking simultaneous measurements of the golf ball motion using two laser vibrometers [38].

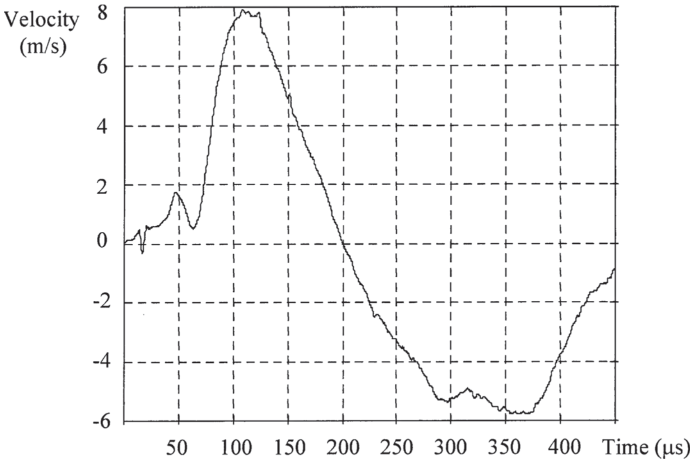


FIGURE 4.20 Golf ball velocity of deformation. (After Hocknell et al. [36].)

The maximum working distance from the laser vibrometer to the target surface is dictated by the need to backscatter sufficient light intensity to the instrument to ensure adequate signal-to-noise ratio for the Doppler signal processor. The latter will often operate successfully at low signal-to-noise ratios whilst offering a reduced frequency response, and there is a trade-off to be obtained here depending upon the measurement requirement. Intensity of backscattered light depends upon the laser power, transmission/collection optics and the scattering properties of the surface as described in Section 4.4.4. Commercially available instruments [30,31] specify working distances of up to 30 m from untreated surfaces, with a laser beam power of less than 1 mW, depending upon the choice of optical head for the vibrometer. This distance can be extended to 200 m if surface treatment is used. Civil engineering applications [39] have seen laser vibrometers used at distances of more than 500 m. At these distances, the speckle on the photodetector may begin to wander due to atmospheric turbulence effects and special speckle-tracking systems are necessary to overcome this problem.

It is important to note that a laser vibrometer measures the 'instantaneous' rate of change in optical length between the instrument and illuminated scattering elements on the target surface (see Equation (4.3)). This is best discussed with reference to Figure 4.21 which shows a vertical laser beam from a laser vibrometer incident on the sloping surface of a wedge target which can move on a horizontally flat surface. With the wedge stationary, if the instrument itself is moved so as to have a non-zero vibration velocity component in the direction of the laser beam (e.g. held-held, see Section 4.4.1), then the optical path length will change and this motion will appear as target vibration. In practice, therefore, it is important to ensure that the vibrometer is effectively vibration-isolated from the target surface over the frequency bandwidth of interest. An accelerometer fixed to the front of the instrument can be used to check the level of this noise source during a movement.

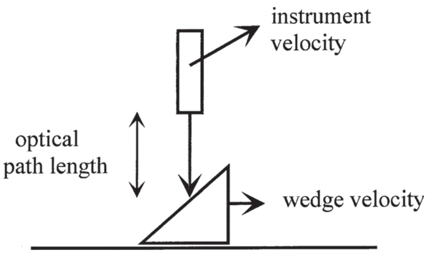


FIGURE 4.21 Optical path length considerations.

A further subtle consideration is the effect on the vibrometer output as the wedge moves on the horizontal surface. Despite the optical path length between the instrument and the target changing, *different* scattering elements are illuminated. The laser vibrometer does not measure the velocity of the wedge, which is in a direction perpendicular to the incident laser beam. Instead, speckle phase noise is produced as scattering elements on the surface of the wedge enter and leave the incident laser beam causing the speckle pattern on the surface of the detector to 'boil' (see Section 4.4.1). It is interesting to note that, if the wedge oscillates, the vibrometer output spectrum will form a periodogram with a fundamental at the oscillation frequency, and this is a classic case of a pseudo-vibration spectrum [24]. For true vibration velocity to be measured, the target surface scatterers must have an 'instantaneous' velocity component in the direction of the incident laser beam when they are in the laser beam.

Problems of speckle phase noise can be alleviated, in practice, by arranging to take a Lagrangian-type measurement where the laser beam is scanned so as to track the same scattering elements during a measurement [40,41]. This approach is analogous to the use of image de-rotators in holographic applications [42].

Laser technology in the engineering workplace is still relatively new, and if its use is to become commonplace, then safety and user friendliness are prime concerns. With current codes of practice for safety, this means that the output beam should be class II, which ensures safety for momentary accidental viewing of the beam and corresponds to an output power of 1 mW or less from a He-Ne laser.

4.5 Measurement of Angular Vibration Velocity

4.5.1 Introduction

Power transmission by machinery is most often achieved through rotating shafts and/or components. Rotating machinery designers have to be crucially aware of the magnitude of angular oscillations and/or torsional resonances inherent in these systems in order to avoid such phenomena as fatigue failure of components, rapid bearing wear, gear hammer, fan belt slippage, excessive noise, etc. A typical problem is the control of torsional resonances of a crankshaft in an internal combustion engine. Engine designers must avoid running the engine at 'critical' speeds where the piston firing frequency coincides with a torsional resonance of the crankshaft system. In higher power engines, it is a standard practice to fit torsional dampers to the crankshaft in order to suppress torsional oscillations over the working speed range of the engine. Damper failure can result in catastrophic fatigue failure of the crankshaft.

Until the advent of laser technology, the measurement of torsional vibration was a particularly difficult measurement problem. Strain gauges with associated telemetry or slip ring systems were used but these are notoriously difficult to fix, calibrate and use successfully. For automotive engine test cell applications, a slotted disc system was often used. The disc, with slots in its periphery, was fixed to the end of the engine crankshaft and a proximity transducer monitored the slot passing frequency as the disc rotated. De-modulation of the slot-passing frequency produced a voltage analogue of the rotational speed and, hence, torsional oscillation within a limited frequency range dictated by the number of slots. In general, the measurement of angular or torsional vibration was problematic for contacting transducer technology and, of course, necessitated machinery downtime whilst arrangements were made for fitting and calibration, etc. The cost of machinery downtime could preclude a measurement being attempted even though the engineer had concluded that a level of torsional oscillation was the root cause of machine failure.

There was, therefore, a real need for a user-friendly instrument which could provide torsional vibration data *in situ*. The advent of laser technology solved this problem, and in what follows, the laser torsional or rotational vibrometer which is now used routinely for this form of measurement is described.

4.5.2 The Laser Torsional or Rotational Vibrometer

The original use of laser technology for the measurement of angular motion was the application of the cross-beam or in-plane vibrometer to measure the tangential surface velocity of a shaft of circular cross section [22]. With reference to Figure 4.13, this instrument could be used when the solid surface remained in the intersection region of the laser beams throughout the course of the measurement. This is commensurate with the shaft or component having a circular cross section and the absence of gross solid body oscillations. In addition, however, the cross-beam vibrometer, in measuring tangential surface velocity, also measures the solid body vibration velocity component which adds to the tangential velocity. It does not distinguish the true angular variation in speed. Successful measurements could, therefore, only be taken where it could be assumed that this effect was negligible.

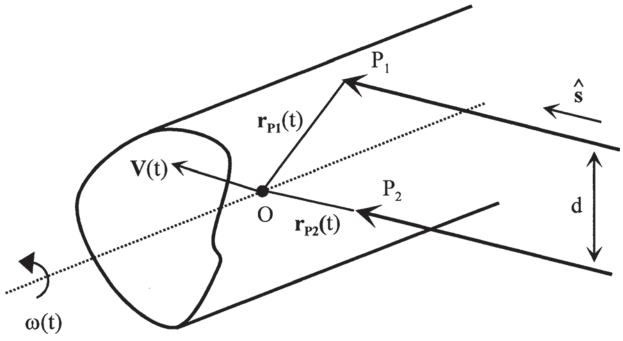


FIGURE 4.22 Laser rotational vibrometer: Theory of operation. (After Halliwell [43,44].)

These problems were overcome using a parallel-beam optical geometry proposed by the author [43,44]. Figure 4.22 shows two parallel laser beams incident on a solid body of arbitrary cross section rotating with an angular velocity $\omega(t)$ about an axis within the body. Without loss of generality, the point O defines an origin on the axis which itself is vibrating with a translational velocity $\boldsymbol{V}(t)$. The two laser beams are incident on the body at points P_1 and P_2 and their direction is defined by the unit vector $\hat{\boldsymbol{s}}$. The position vector of the points P_1 and P_2 with respect to the origin are $\boldsymbol{r}_{p1}(t)$ and $\boldsymbol{r}_{p2}(t)$, respectively. We now consider the total velocity of the scattering elements $\boldsymbol{V}_{p1}(t)$ and $\boldsymbol{V}_{p2}(t)$ illuminated by the laser beams at P_1 and P_2, respectively, which can be written

$$\boldsymbol{V}_{p1}(t) = \omega(t) \times \boldsymbol{r}_{p1} + \boldsymbol{V}(t) \tag{4.12}$$

$$\boldsymbol{V}_{p2}(t) = \omega(t) \times \boldsymbol{r}_{p2} + \boldsymbol{V}(t) \tag{4.13}$$

With reference to Equation (4.2), the Doppler shift f_{DP_1} and f_{DP_2} in light directly backscattered from P_1 and P_2, respectively, is given by

$$f_{DP_1} = \frac{2}{\lambda}\hat{\boldsymbol{s}}\cdot\boldsymbol{V}_{P_1}(t) \qquad f_{DP_2} = \frac{2}{\lambda}\hat{\boldsymbol{s}}\cdot\boldsymbol{V}_{P_2}(t). \tag{4.14}$$

Mixing this backscattered light onto a photodetector produces a beat frequency f_B in the output [3] proportional to the difference $f_{DP_1} - f_{DP_2}$ which substituting for $\boldsymbol{V}_{P_1}(t)$ and $\boldsymbol{V}_{P_2}(t)$ from Equations (4.12) and (4.13), respectively, is given by

$$f_B = \hat{\boldsymbol{s}}\cdot[\boldsymbol{\omega}(t)\times(\boldsymbol{r}_{P_1}(t)-\boldsymbol{r}_{P_2}(t)] = \boldsymbol{\omega}(t)\cdot[(\boldsymbol{r}_{P_1}(t)-\boldsymbol{r}_{P_2}(t))\times\hat{\boldsymbol{s}}]. \tag{4.15}$$

This result shows that the beat frequency is insensitive to the translational vibration velocity $\boldsymbol{V}(t)$. The modulus of the vector product in Equation (4.15) is the separation distance d of the laser beams. Substituting for the angular velocity vector and taking the scalar product [45] show that

$$f_B = \left(\frac{4\pi}{\lambda}\right)Nd\sin\theta \tag{4.16}$$

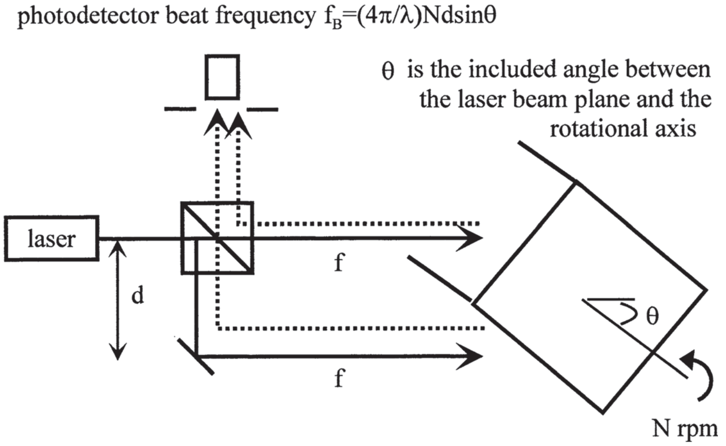

photodetector beat frequency $f_B=(4\pi/\lambda)Nd\sin\theta$
θ is the included angle between the laser beam plane and the rotational axis
laser
d
f
f
θ
N rpm

FIGURE 4.23 Angular velocity measurements. (After Halliwell [45].)

where N is the rotational speed of the body and θ is the included angle between the rotational axis of the body and the plane defined by the incident laser beams. In this way, frequency demodulation of the photodetector signal provides a time-resolved voltage analogue of the speed of the rotating body $N(t)$, the ac part of which is the angular oscillation or torsional vibration.

The optical geometry proposed by the author to produce the beat frequency is shown in Figure 4.23. This parallel beam geometry is now used in commercially available laser torsional or rotational vibrometers used for the measurement of angular velocity.

It is important to note that Equation (4.16) also shows that the measurement is not sensitive to the cross-sectional shape of the body. This fact allows a laser rotational vibrometer to work successfully on targets of arbitrary cross section. The beat frequency is simply dependent upon the rotational speed N and the laser beam separation d. The sin θ term shows that the output of a laser rotational vibrometer is proportional to the component of angular velocity of the body which is resolved along the direction of the normal to the laser beam plane. Discussions of the effect of variations in θ are included in Section 4.5.4.

With reference to Equation (4.16), successful operation of a vibrometer which uses direct optical heterodyning of scattered light in this way relies on a non-zero mean value of the rotational speed N. This means that the vibrometer is unable to measure the target 'tilt' or angular oscillation without full rotations. A solution is to frequency-shift one of the incident laser beams relative to its parallel counterpart [46]. In this way, a stationary target still provides a 'carrier' beat frequency in the photodetector output which the subsequent angular velocity of the target frequency modulates.

Direct optical heterodyning of the scattered light from the two incident laser beams means that two laser speckle patterns, which repeat with each rotation of the target, are continually interfering on the photodetector surface. The beat frequency signal-to-noise ratio is inherently lower in this situation compared with that obtained by the interference of a single speckle pattern with a frequency-shifted reference beam as per Equation (4.7). With reference to this equation, this is because the reference beam intensity term I_R is maintained at a constant high level.

Lewin et al. [47] proposed an alternative interferometer arrangement which avoids direct optical heterodyning of the scattered light and allows for measurement of angular oscillation without the need for a target non-zero mean rotational speed. This arrangement is shown in Figure 4.24. With reference to this figure, two separate heterodyne interferometers (see Figure 4.4) are used to provide the parallel beam arrangement. Each interferometer is frequency-shifted by the same Bragg cell and both share the same laser source. The beat frequency, proportional to rotational speed (as per Equation (4.16)), is obtained by electronically mixing the photodetector outputs from each interferometer. In this way, the two significant advantages of a higher signal-to-noise operation and detection of angular oscillation in the presence of zero mean rotational speed are obtained.

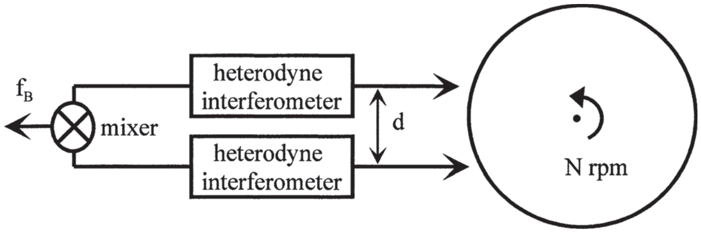


FIGURE 4.24 Angular velocity measurement. (After Lewin et al. [47].)

The use of separate optical interferometers in laser rotational vibrometers has been adopted by commercial instrument manufacturers [48]. Laser rotational vibrometers are commercially available with specifications that include speed ranges in excess of 10 000 rev min^{-1}, frequency response in excess of 1 kHz. The accuracy figure quoted for torsional vibration amplitude (ac rotational vibration) is approximately less than 2% of the full-scale reading.

4.5.3 Measurements within a Rotating System

It is important to note that a laser rotational vibrometer measures the instantaneous component of *total* angular velocity of a body resolved along the normal to the laser beam plane. Figure 4.25 shows a target, of arbitrary cross-sectional shape, rotating with angular velocity Ω_2 about an axis passing through a point O′ in space. With reference to this figure, the target is also rotating with angular velocity Ω_1 about an axis through an origin O which, without loss of generality, is fixed in space. Such a situation may be envisaged, in practice, by a rotating target fixed to the end of a robot arm which is itself undergoing angular motion. For simplicity, the analysis is confined to one laser beam, incident at a point P on the target, and without loss of generality, solid body translation motion of the target is ignored.

The total velocity $\boldsymbol{V}_{\mathrm{P}}$ of the scattering elements illuminated by the laser beam is given by

$$\boldsymbol{V}_{\mathrm{P}} = \boldsymbol{\Omega}_1 \times \boldsymbol{r}_1 \times \boldsymbol{\Omega}_2 \times \boldsymbol{r}_2 \tag{4.17}$$

where $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are the distances of P from the points O and O′, respectively.

The latter are defined by the equations

$$\boldsymbol{r}_1 = \boldsymbol{a}_1 + p(t)\hat{\boldsymbol{s}} \qquad \boldsymbol{r}_2 = \boldsymbol{a}_2 + p(t)\hat{\boldsymbol{s}} \tag{4.18}$$

where $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are the distances from O and O′ to a *fixed point* A on the laser beam axis, respectively. The unit vector $\hat{s}$ defines the laser beam direction and $p(t)$ is a time-varying scalar which defines the distance from A to the target surface at P. Variations in $p(t)$ are dictated by the shape of the target surface as the latter rotates about both axes.

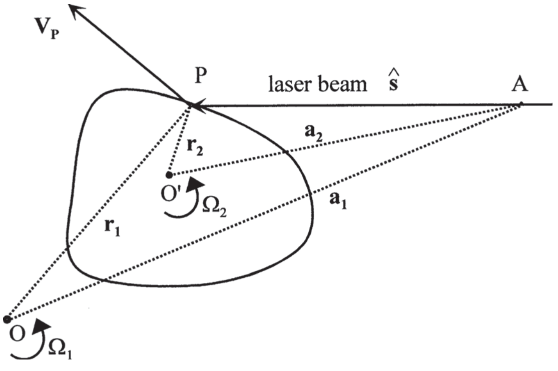


FIGURE 4.25 Total angular velocity measurement.

The Doppler shift in backscattered light f_{P} from P (see Equation (4.2)) is given by

$$f_{\mathrm{P}} = \frac{2}{\lambda}\hat{\boldsymbol{s}} \cdot \boldsymbol{V}_{\mathrm{P}}. \tag{4.19}$$

Substituting for $\boldsymbol{V}_{\mathrm{P}}$, $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ from Equations (4.17) and (4.18), respectively, the triple scalar product terms $\hat{\boldsymbol{s}} \cdot (\boldsymbol{\Omega}_1 \times p(t)\hat{\boldsymbol{s}})$ and $\hat{\boldsymbol{s}} \cdot (\boldsymbol{\Omega}_2 \times p(t)\hat{\boldsymbol{s}})$ are identically zero to give

$$f_{\mathrm{P}} = \frac{2}{\lambda}[\boldsymbol{\Omega}_1 \cdot (\boldsymbol{a}_1 \times \hat{\boldsymbol{s}}) + \boldsymbol{\Omega}_2 \cdot (\boldsymbol{a}_2 \times \hat{\boldsymbol{s}})]. \tag{4.20}$$

This result is analogous to that demonstrated by Equation (4.10) in that the Doppler shift is independent of the cross-sectional shape of the target. When the second parallel laser beam from the laser rotational vibrometer is considered, an equivalent result for the Doppler shift in backscattered light $f_{\mathrm{P'}}$ from P′ produces

$$f_{\mathrm{P'}} = \frac{2}{\lambda}[\boldsymbol{\Omega}_1 \cdot (\boldsymbol{a}_1' \times \hat{\boldsymbol{s}}) + \boldsymbol{\Omega}_2 \cdot (\boldsymbol{a}_2' \times \hat{\boldsymbol{s}})]. \tag{4.21}$$

After heterodyning, the beat frequency f_{B} is proportional to the difference $f_{\mathrm{P}} - f_{\mathrm{P'}}$ given by

$$f_{\mathrm{B}} = \frac{2}{\lambda}[\boldsymbol{\Omega}_1 \cdot ([\boldsymbol{a}_1 - \boldsymbol{a}_1'] \times \hat{\boldsymbol{s}}) + \boldsymbol{\Omega}_2 \cdot ([\boldsymbol{a}_1 - \boldsymbol{a}_2'] \times \hat{\boldsymbol{s}})]. \tag{4.22}$$

The modulus of the vector products in Equation (4.22) is identically equal to the laser beam separation distance d. The final result, substituting for the angular velocity vectors, analogous to that presented in Equation (4.16), is given by

$$f_{\mathrm{B}} = \frac{2}{\lambda}[N_1 d \sin\theta + N_2 d \sin\alpha] \tag{4.23}$$

where θ and α are the included angles between the laser beam plane and the rotational axes defined by $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$, respectively. N_1 and N_2 are the rotational speeds in Hz about these axes, respectively.

In this way, it is important to note that the output of a laser rotational vibrometer is proportional to the modulus of the *total* component of angular velocity of the target, resolved along the normal to the incident laser beam plane.

4.5.4 Practical Considerations and Examples of Use

Obtaining the calibration of voltage output per revolution from a laser rotational vibrometer can usually be carried out *in situ*. It is not necessary to attempt to measure the beam separation d or the angle θ between the laser beam plane and the rotational axis. In most practical situations, it is possible to rotate the target at several known speeds so that an *in situ* calibration of dc volts per revolution is possible. If this is not the case, then a small disc can be used which is driven at known speeds and interposed, in the laser beams, in place of the target. It is straightforward to monitor the dc voltage from the instrument at three known speeds to produce a calibration factor giving negligible error. It is usual to use the vibrometer in conjunction with a real-time spectrum analyzer which will integrate the output directly to produce vibration spectra in terms of torsional vibration or angular oscillation displacement referenced to degrees peak. If a real-time analyzer is used *in situ* a measure of the fundamental frequency of the periodogram produced by the speckle noise (see Section 4.4.1) can be conveniently noted to provide an alternative measure of the rotational speed of the target for calibration purposes. If a real-time analyzer is not used the vibrometer output can be tape recorded for later analysis, but it must be remembered that the dc signal level is needed if an *in situ* calibration has taken place.

In practice, it is often more convenient to tripod-mount the instrument so that the plane of the laser beams is incident at an arbitrary angle to the rotational axis of the target (e.g. the end face of a rotating shaft). It should be noted, however, that the dependency of the beat frequency f_B on this angle, defined by Equation (4.16), means that a local, time-varying, tilt of the axis will modulate f_B and contribute a source of noise. Miles et al. [49] have examined this effect in detail and concluded that the error is negligible for values of $70° \leq \theta \leq 90°$. The ideal mode of operation for minimum error is for $\theta = 90°$ corresponding to the plane of the incident beams being at right angles to the local rotational axis. It is useful to remember that if the instrument is hand-held, tilt of the operator will only affect the spectra obtained below a frequency of 30 Hz [25].

Laser rotational vibrometers are a convenient means of measuring the revolution rate of a rotating target. The more usual measurand of interest, however, is the angular oscillation or torsional vibration spectrum. In practice, the highest levels of torsional vibration can be anticipated on the crankshafts of large diesel engines or reciprocating compressors. These levels may be as high as a few degrees of peak displacement, but in Doppler frequency terms, these levels still represent a small modulation of the beat frequency f_B. Consequently, measuring a high level of torsional vibration does not usually present a problem and it is necessary to ensure simply that the levels are within the dynamic range of the Doppler signal processor.

The lowest level of torsional vibration which can be measured is limited by speckle dynamics, as discussed in Section 4.4.1. A rotating (or tilting) target produces speckle noise in the spectrum which forms a periodogram at the fundamental rotation frequency. This is true whether a direct optical heterodyne or separate optical interferometers have been used to provide the beat frequency. Figure 4.26 shows a displacement spectrum obtained from a measurement of the speed of rotation of a dc motor in a tape recorder. The spectrum exhibits the periodic nature of the speckle noise which completely masks the intended speed variation measurement. Fortunately, the insensitivity of the geometry to solid body translational oscillation means that the speckle periodicity can be removed by moving the laser beam plane during the course of a measurement (whilst avoiding tilt of the plane!) so that the speckle patterns incident on the photodetectors do not repeat with each revolution. Figure 4.27 shows an equivalent speed measurement of the tape-recorder motor when the incident laser beam plane was moved from side to side at a frequency of approximately 1 Hz. The figure shows that the speckle noise peaks have been effectively removed and the 'wow' of the motor can be distinguished. For most mechanical engineering applications, a level of −60 dB re 1° peak means that torsional vibrations are negligible and speckle noise periodicity can be ignored.

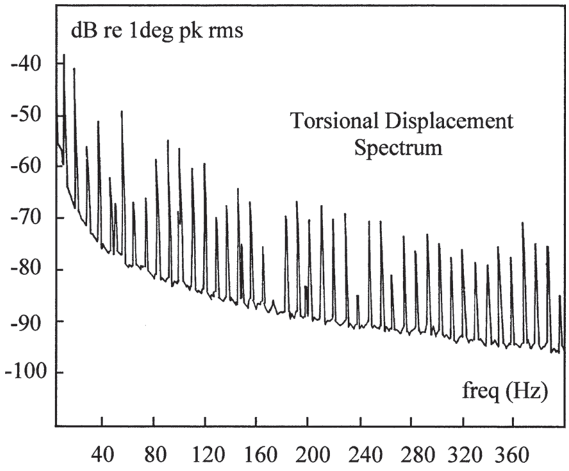


FIGURE 4.26 Periodic speckle noise.

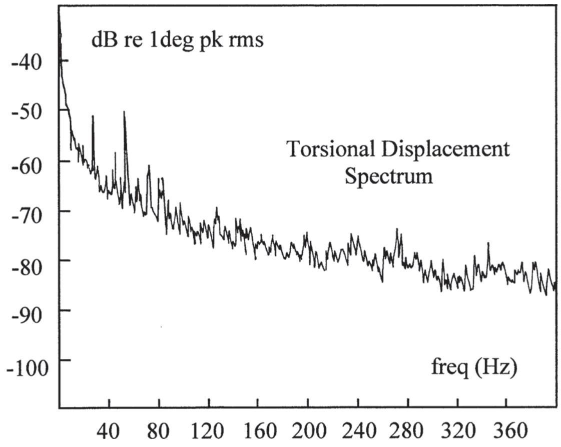


FIGURE 4.27 Speckle noise peak reduction.

The range of possible rotational speeds, found in practice, presents a problem for Doppler signal processors. At a beam separation d of 10 mm, for a wavelength $\lambda = 0.6\,\mu m$ and a beam plane perpendicular to the rotational axis ($\theta = 90°$), a speed variation of 500–15 000 rev min^{-1} produces a beat frequency range of 1.5–45 MHz. The theory analyzing the Doppler shift in the backscattered light [45] assumes that the incident laser beams are infinitely thin, and therefore, by an order of magnitude argument, a minimum beam separation of typically 10 mm is required. Commercial instruments usually have a fixed beam separation of this order and employ different speed ranges to address this problem. At higher rotational speeds, use of a vibrometer so that the incident laser beam plane is at an angle to the rotational axis can reduce the mean beat frequency, but attention must be paid to the problem of error caused by local tilt of the rotational axis during a measurement, as discussed earlier [49].

Very low values of rotational speed do not cause a problem for rotational vibrometers which utilize frequency-shifting and separate interferometers to provide the parallel beam geometry. This is because frequency-shifting inherently provides the necessary carrier frequency for a Doppler signal processor independent of the magnitude of the rotational speed. At high rotational speeds, however, care must be taken that the Doppler signal produced by a separate interferometer remains within the dynamic range of the signal processor, e.g. for a laser beam of wavelength $\lambda = 0.6\,\mu\text{m}$, incident in a direction perpendicular to the rotational axis of a shaft, at a distance of 50 mm from the latter, a rotational speed of 5000 rev min^{-1} produces a Doppler shift in the backscattered light of 87 MHz! This consideration may prevent the use of a vibrometer in situations where a measurement cannot be taken with the incident laser beams positioned on either side of or close to the rotational axis of the target.

REFERENCES

1. Shaw S 1982 Noise control of stone cutting circular saws *MSc Thesis* ISVR, University of Southampton, UK.
2. Dainty C J (ed) 1984 *Laser Speckle and Related Phenomena* (Berlin: Spinger).
3. Watrasiewicz B M and Rudd M J 1976 *Laser Doppler Measurements* (London: Butterworths).
4. Ballantyne A, Blackmore C and Rizzo J E 1974 Frequency shifting for laser anemometers by scattering *Opt. Laser Technol.* **6** 170–3.
5. Drain L E and Moss B C 1972 The frequency shifting of laser light by electro-optic techniques *Opto-Electron.* **4** 429.
6. Oldengarm J, van Krieken A H and Raternik H J 1973 Laser Doppler velocimeter with optical frequency shifting *Opt. Laser Technol.* **5** 249–52.
7. Baker J R, Laming R I, Wilmshurst T H and Halliwell N A 1990 A new high sensitivity laser vibrometer *Opt. Laser Technol.* **22** 241–4.
8. Lewin A C, Kersey A D and Jackson D A 1985 Non-contact surface vibration analysis using a monomode fibre optic interferometer incorporating an open air path *J. Phys. E: Sci. Instrum.* **18** 604–7.
9. Buchave P 1975 Laser Doppler vibration measurement using variable frequency shift *DISA Inf. Bull.* **18** 15–29.
10. Laming R I, Gold M P, Payne D N and Halliwell N A 1985 Fibre-optic vibration probe *Electron. Lett.* **22** 167–8.
11. Halliwell N A 1979 Laser Doppler measurement of vibration surfaces: a portable instrument *J. Sound Vibration* **62** 312–15.
12. Pickering C J D and Halliwell N A 1986 The laser vibrometer: a portable instrument *J. Sound Vibration* **107** 471–85.
13. Bruel and Kjaer Type 3544, laser velocity transducer, B and K Ltd, Denmark.
14. Oldengarm J 1977 Development of rotating diffraction gratings and their use in laser anemometry *Opt. Laser Technol.* April 69–71.
15. Scruby C B and Drain L E 1990 *Laser Ultrasonics* (Bristol: Adam Hilger) pp 116–22.
16. Reibold R and Molkenstruck W 1981 *Acoustica* **49** 205.
17. Ometron Ltd 1997 VPI 4000 User Manual, Ometron, London.
18. Selbach H and Lewin A C 1987 Differentielles fiberoptisches laser Doppler Vibrometer zur Schwingungsanalyse *Laser 1987—Optoelectronics in Engineering (Munich).*
19. Polytec GmbH, D-76337, Waldbronn, Germany.
20. Laming R I, Wilmshurst T H, Halliwell N A and Baker J R 1990 A practical all-fibre laser vibrometer *Exp. Techniques* March/April 44–7.
21. Drain L E 1980 *The Laser Doppler Technique* (Chichester: Wiley).
22. Halliwell N A, Pullen L and Baker J 1983 Diesel engine health: laser diagnostics, SAE technical paper series 831324 *International Off-highway Meeting and Exposition (Milwaukee) also in Trans. SAE* 986–94.
23. Wilmshurst T H 1974 An autodyne frequency tracker for use in laser Doppler anemometry *J. Phys. E: Sci. Instrum.* **7** 924.
24. Rothberg S J, Baker J R and Halliwell N A 1989 Laser vibrometry: pseudo vibrations *J. Sound Vibration* **135** 516–22.
25. Griffin M J 1990 *Handbook of Human Vibration* (London: Academic).
26. Rothberg S J and Halliwell N A 1994 Vibration measurements on rotating machinery using laser Doppler velocimetry *J. Vibration Acoust.* **116** 326–31.
27. Rothberg S J and Halliwell N A 1995 Application of laser vibrometry to vibration measurement on rotating components *Proc. 15th ASME Biennial Conf. on Mechanical Vibration and Noise* DE 84–3 Part C, pp 1425–34.
28. Bell J R and Rothberg S J 1998 Radial vibration measurements on rotors using laser vibrometry: a first practical solution to the cross-sensitivity problem *Proc. SPIE 3rd Int. Conf. on Vibration Measurements by Laser Techniques: Advances and Applications* **3411** 14–22.
29. Bell J R and Rothberg S J 1999 Laser vibrometry: the complete six degree-of-freedom model of target velocity sensitivity ASME *Proc. Design Eng. Tech. Conf: Mechanical Vibration and Noise.*
30. Laser Vibrometers, Polytec GmbH, W-7517, Waldbronn, Germany.
31. VPI Sensor, Ometron Ltd, London, UK.
32. Chowdhury M R and Hall R L 1998 Vibration measurements of Olmsted prototype wickets by laser techniques *Proc. SPIE 3rd Int. Conf. on Vibration Measurements by Laser Techniques: Advances and Applications* **3411** 266–74.
33. Rossi G L, Santolini M, Giachi M and Generosi S 1996 Dynamic characterization of a centrifugal compressor rotor by a laser scanning vibrometer *Proc. SPIE 2nd Int. Conf. on Vibration Measurements by Laser Techniques: Advances and Applications* **2868** 290–9.
34. Stanbridge A B and Ewins D J 1995 Modal testing of rotating discs using a scanning LDV *Proc. ASME Design Engineering Technical Conf.* (Boston, MA) DE–Vol 84–2 Vol 3B

35 Sanbridge A B and Ewins D J 1996 Using a continuously scanning laser Doppler vibrometer for modal testing *Proc. Int. Modal Analysis Conf.* (Dearborn) **14**.

36. Hocknell A, Jones R and Rothberg S J 1996 Experimental analysis of impacts with large elastic deformation: 1. Linear motion *Meas. Sci. Technol.* **7** 1247–54.

37. Halliwell N A, Kember S A and Rothberg S J 1998 Laser vibrometry for impact measurements *Proc. Euromech. Colloquium 386, Dynamics of Vibro-Impact Systems* ed V I Babitsky (Berlin: Springer) pp 279–88.

38. Hocknell A, Jones R and Rothberg S J 1998 Remote vibration measurements: compensation of waveform distortion due to whole body translations *J. Sound Vibration* **214** 285–307.

39. Bougard A J and Ellis B R 1998 Laser measurement of building vibration and displacement *Proc. SPIE 3rd Int. Conf. on Vibration Measurements by Laser Techniques: Advances and Applications* **3411** 254–65.

40. Bucher I, Schmiechen P, Robb D A and Ewins D J 1994 A laser-based measurement system for measuring the vibration on rotating discs *Proc. SPIE 1st Int. Conf. on Vibration Measurements by Laser Techniques: Advances and Applications* **2358** 398–408.

41. Castellini P and Santolini C 1996 Vibration measurements on blades of naval propellers rotating in water *Proc. SPIE 2nd Int. Conf. on Vibration Measurements by Laser Techniques: Advances and Applications* **2868** 186–94.

42. Fieldhouse J D and Newcomb T P 1996 Double pulsed holography used to investigate noisy brakes *Opt. Lasers Eng.* **25** 455–595.

43. Halliwell N A, Pickering C J D and Eastwood P G 1984 The laser torsional vibrometer: a new instrument *J. Sound Vibration* **93** 588–92.

44. Halliwell N A and Eastwood P G 1985 The laser torsional vibrometer *J. Sound Vibration* **101** 446–9.

45. Halliwell N A 1996 The laser torsional vibrometer: a step forward in rotating machinery diagnostics *J. Sound Vibration* **190** 399–418.

46. Halliwell N A, Hocknell A and Rothberg S J 1997 On the measurement of angular vibration displacements: a laser tiltmeter *J. Sound Vibration* **208** 497–500.

47. Lewin A C, Roth V and Siegmund G 1994 New concept for interferometric measurement of rotational vibrations *Proc. SPIE 1st Int. Conf. on Vibration Measurements by Laser Techniques: Advances and Applications* **2358** 24–36.

48. Polytec, Rotational Vibrometer Series 4000 Polytec Ltd, Waldbronn, Germany.

49. Miles T J, Lucas M, Halliwell N A and Rothberg S J 1999 Torsional and bending vibration measurements on rotors using laser technology *J. Sound Vibration* **226** 441–67.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5

Electronic Speckle Pattern Interferometry (ESPI)

Lianxiang Yang and Xinya Gao

CONTENTS

5.1 Introduction ... 61
5.2 Principle of ESPI ... 62
  5.2.1 Introduction ... 62
  5.2.2 Description of the Technique ... 62
  5.2.3 Principle of Fringe Formation ... 63
  5.2.4 Fringe Interpretation: Relationship between the Phase Change $\Delta(x,y)$ and Deformation Components *u*, *v*, and *w* ... 63
  5.2.5 Measurement of Out-of-Plane and In-Plane Deformation ... 64
5.3 Evaluation of the Interference Phase in ESPI ... 65
  5.3.1 Temporal Phase-Shift Technique ... 65
    5.3.1.1 4+4 Algorithm ... 65
    5.3.1.2 4+1 Algorithm ... 66
    5.3.1.3 Other Temporal Phase-Shift Algorithms ... 66
    5.3.1.4 Generation of a Phase-Shift ... 67
  5.3.2 Spatial Phase-Shift Technique ... 67
    5.3.2.1 Spatial Phase-Shift Method Based on Multi-Pixel Calculation ... 67
    5.3.2.2 Spatial Phase-Shift Method Based on Carrier-Frequency and Fourier Transformation ... 68
5.4 Applications of ESPI ... 70
  5.4.1 Applications of Temporal Phase-Shift ESPI ... 70
    5.4.1.1 Out-of-Plane and In-Plane Deformation Measurement and Non-destructive-Test ... 70
    5.4.1.2 Measurement of Three-Dimensional Deformations and Strain ... 71
  5.4.2 Applications of Spatial Phase-Shift ESPI ... 73
    5.4.2.1 Out-of-Plane and In-Plane Deformation Measurement under a Continuous Loading ... 73
    5.4.2.2 Simultaneous Measurement of 3D-Deformations under a Single Loading ... 74
  5.4.3 Applications for Dynamic Measurements ... 75
    5.4.3.1 Time-Averaged Method with a Refreshed Reference Frame ... 75
    5.4.3.2 Stroboscopic Method ... 76
    5.4.3.3 Double-Pulse Method ... 77
    5.4.3.4 Further applications of ESPI ... 77
5.5 Conclusions ... 77
References ... 78

5.1 Introduction

Speckle metrology includes techniques such as speckle photography, photography-based speckle interferometry, electronic speckle pattern interferometry (ESPI) and speckle shearography. The most important technique of speckle metrology for engineering measurements is the ESPI, also currently called digital speckle pattern interferometry (DSPI). The purpose of this chapter is to present recent developments and applications of ESPI for measurement of three-dimensional deformations and strain as well as for non-destructive testing and quality inspections. Considering that people engaged in these areas are mainly researchers, engineers, and graduate students in engineering departments who work in engineering measurement or in experimental and photo mechanics, this chapter intends to convey how relatively easily ESPI can be applied by the non-specialist in the field of optics or, alternatively, by researchers in engineering areas such as the authors. In writing this chapter, the authors will attempt to provide a unified, self-contained treatment of the theory, practice and application of ESPI and hope it can serve as a reference for researchers in these areas.

Please note that deformations and strains, both under static and dynamic loading, can also be measured by digital image correlation (DIC), and nowadays, DIC is very active due to its simplicity. However, these two technologies have different measurement ranges. ESPI measures deformation from a few micrometres down to dozens of nanometres, whereas DIC measures deformation in a range from a few micrometres up to a few millimetres. Therefore, these two technologies cannot be replaced with each other. Instead, they complement each other.

5.2 Principle of ESPI

5.2.1 Introduction

ESPI, developed from speckle pattern interferometry (SPI), is a whole field deformation measurement technique that uses laser speckles as the carrier of deformation. The phenomenon of laser speckles was first reported in 1962 by Rigden and Gordon of Bell Laboratories [1]. Laser speckles, with the appearance of randomly distributed light and dark spots, are generated when a highly coherent light is modulated by a diffusely reflecting (or transmitting) object. In the description of holographic methodologies, the speckles were viewed as a disturbance to be suppressed or eliminated because they severely limited the effective resolution. In the early 1970s, Leendertz and Butters first put forward a proposal, “If we cannot get rid of speckle, why don’t we used it?” This suggestion has now assumed considerable significance and played an important role in the development of speckle-measuring techniques in optical metrology. One of the most important methods in the area of speckle metrology is SPI [2].

SPI was developed for measurement of deformation and used to determine strain/stress and vibration [3]. Traditional SPI requires photographic recording, wet development, and optical reconstruction, thus making it very time-consuming. SPI did not actually gain wide acceptance by industry and researchers until electronic processing was applied to it. With electronic processing applied, the technique became known as ESPI [4]. ESPI records the speckle patterns before and after an object is loaded using a charge-coupled device (CCD) camera, then compares and processes the information by electronic methods and displays the result, similarly to interference fringe pattern, on a monitor. Compared with photographic techniques, ESPI eliminates photographic recording, wet processing, and image reconstruction and, thus, realizes a real-time display of the speckle interference fringe pattern. The advantages of this approach are readily apparent.

ESPI enables a real-time observation of the fringe pattern. An interpretation of the fringe pattern remains an undesirable obstacle as it is still at the same technology level as photographic versions, where the measured results can only be evaluated visually. The deformation at a point of interest is generally determined by means of multiplying the system constant by the fringe order at the point. The difficulty stems from the ambiguities in determining fringe orders and their signs. Therefore, some additional information is needed to remove the ambiguities associated with a quantitative determination of fringe orders. At this level of development, the common practice is to identify fringe orders based on prior knowledge of the problem, such as the known boundary conditions and the nature of the problem. This problem has been a major stumbling block for expanding usage in industry.

Automatic and quantitative evaluation of the fringe pattern became possible with the introduction of the phase-shift technique [5] in ESPI as reported by Creath. Since that time, different algorithms have been developed and used for phase calculation and determination. ESPI implemented by the phase-shift technique has been considered as the second generation of ESPI. They are also known by the name of phase-shifting SPI as well as DSPI [6] by some other authors. With continuous developments and improvements in software programs, computer technology and optical measuring techniques, the applications of ESPI are growing in many areas from automotive and aeronautic/astronautic to high-tech and biomedical industries [7,8]. In this chapter, the principles of ESPI will be briefly reviewed. Their recent developments are presented, and their applications will be demonstrated by examples of deformation and strain measurement, vibration analysis and Non destructive Testing (NDT).

5.2.2 Description of the Technique

The principle of an ESPI system is illustrated in Figure 5.1. A laser is divided by a beam splitter into two beams: the object beam in the illumination direction and the reference beam. The object beam illuminates a rough surface of a measured object after it is expanded. Then the scattered light is collected by an imaging lens which forms an image on a CCD chip of a camera. The angle α between the illumination and observation direction is called the illumination angle. The reference beam also strikes the CCD chip through a beam combiner. At the CCD chip, the reference beam interferes with the object beam. As a result, a speckle interferogram, also known as a speckle

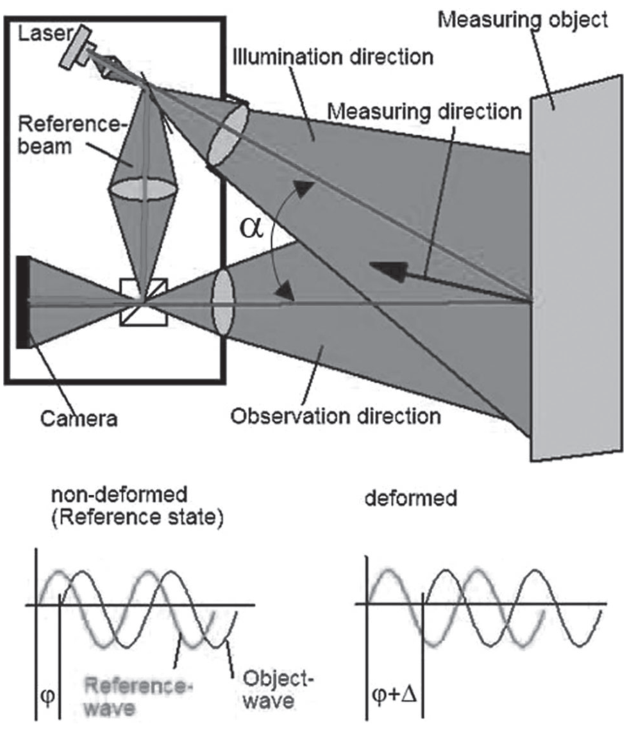


FIGURE 5.1 Principle of the ESPI system.

pattern, is formed. The intensity of the speckle pattern $I(x, y)$ captured by the camera can be expressed as:

$$I(x,y) = A(x,y) + B(x,y)\cos\varphi(x,y) \qquad (5.1)$$

where $A(x, y)$ is the background intensity, $B(x, y)$ is the modulation amplitude of the speckle interferogram, and $\varphi(x, y) = \theta_0(x, y) - \theta_R(x, y)$ is the phase difference between the object beam phase $\theta_0(x, y)$ and the reference beam phase $\theta_0(x, y)$.

5.2.3 Principle of Fringe Formation

Fringes or phase maps display the measured information in most interferometry methods. In the intensity equation of the speckle interferogram described in Equation (5.1), no visible fringes can be observed because the phase difference $\varphi(x, y)$ or the phase distribution $\theta_0(x, y)$ is a high frequency and random term; thus, one can see only a random speckle which is related to object surface roughness. When the object is deformed, the phase difference $\varphi(x, y)$ is changed because the object beam phase is changed to $\theta_0(x, y)$. The intensity of the speckle pattern after loading becomes:

$$I'(x,y) = A(x,y) + B(x,y)\cos\left[\varphi(x,y) + \Delta(x,y)\right] \qquad (5.2)$$

where $\Delta(x, y) = [\theta_0^{'}(x, y) - \theta_R^{'}(x, y)] - [\theta_0(x, y) - \theta_R(x, y)]$, and $\theta_0(x, y)$ and $\theta_R(x, y)$ represent the object beam phase $\theta_0(x, y)$ and the reference beam phase $\theta_R(x, y)$ after deformation, respectively. Because nothing happens to the reference beam during the loading, Δ is only equal to the change of the object beam due to the loading, in which $\Delta(x, y) = \theta_0^{'}(x, y) - \theta_0^{'}(x, y)$, as shown in the bottom of Figure 5.1. Subtracting the Equation (5.2) from (5.1) generates:

$$I_{\text{sub}}(x,y) = \left|I(x,y) - I'(x,y)\right| = B(x,y)\left|\sin[\varphi(x,y) + \frac{\Delta(x,y)}{2}]\sin[\frac{\Delta(x,y)}{2}]\right| \qquad (5.3)$$

Because intensity data cannot be negative, the absolute value is displayed. Equation (5.3) showing the first sin-term, dominated by the high-frequency phase $\varphi(x,y)$, is a high-frequency term. It is modulated with a second sin-term with the object beam phase change $\Delta(x,y)$ as an argument. The phase change $\Delta(x,y)$ is caused by a loading and is usually a low-frequency term. Figure 5.2a is a speckle pattern before loading, which is described by Equation (5.1); Figure 5.2b is a speckle pattern after loading, as described by Equation (5.2); Figure 5.2c shows fringes due to the loading obtained by subtraction of the two intensity equations as shown in Equation (5.3). Dark fringes in Figure 5.2c (where $I_{\text{sub}}(x,y) = 0$) can be observed if $\Delta(x,y)$ is equal to $N \times 2\pi$, $N = 0, \pm 1, \pm 2, \pm 3, \ldots$, where N is the fringe order). By visual observation of the fringe order from the resulting image, the phase change $\Delta(x,y)$ can be determined at the positions of the fringe orders.

5.2.4 Fringe Interpretation: Relationship between the Phase Change $\Delta(x,y)$ and Deformation Components *u*, *v*, and *w*

As described in Section 5.2.3, the fringe at an arbitrary point of the speckle pattern depicts a phase change $\Delta(x,y)$ of the object beam as introduced by a deformation. Therefore, the phase change $\Delta(x,y)$ is related to the deformation components u, v and w as follows [4,6,9]:

$$\Delta = \frac{2\pi}{\lambda}(\mathrm{A}u + \mathrm{B}v + \mathrm{C}w) \qquad (5.4)$$

where λ is the wavelength of the laser used, and A, B and C are sensitivity factors related to the locations of the camera and the laser source. If the object size is much smaller than the distances from the object to the laser source and the camera, the sensitivity factors A, B and C can be simplified as $\mathrm{A} = \sin\alpha$, $\mathrm{B} \approx 0$, and $\mathrm{C} = (1 + \cos\alpha)$, and Equation (5.4) can be simplified as follows [10]:

$$\Delta(x,y) = \frac{2\pi}{\lambda}[u(x,y)\sin\alpha + w(x,y)(1 + \cos\alpha)] \quad (\text{if } \alpha \text{ lies in xoz plane}) \qquad (5.5)$$

where α is the illumination angle between the illumination and the observation directions as shown in Figure 5.1. Similarly, when the illumination angle is in the YOZ plane, Equation (5.5) becomes:

$$\Delta(x,y) = \frac{2\pi}{\lambda}[v(x,y)\sin\alpha + w(x,y)(1 + \cos\alpha)] \quad (\text{if } \alpha \text{ lies in yoz plane}) \qquad (5.6)$$

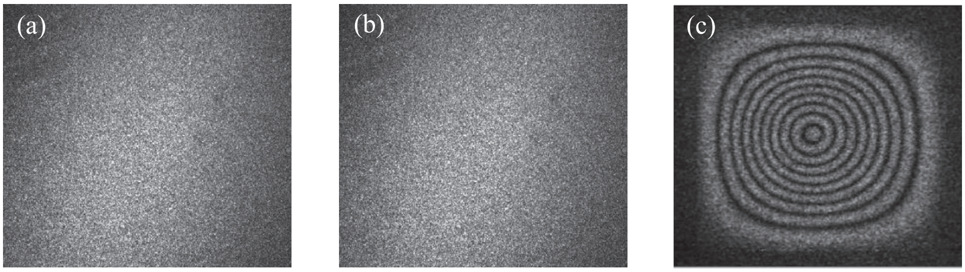


FIGURE 5.2 Fringe formation: (a) a speckle pattern before loading; (b) the speckle pattern after loading; (c) a fringe pattern obtained by digital subtraction of (b) from (a) for a square plate clamed all around and loaded centrally.

Equations (5.5) and (5.6) represent the fundamental equations of ESPI. In these equations, the phase change $\Delta(x,y)$ can be calculated based on counting the fringes. However, the deformation components u and w in Equation (5.5) or v and w in Equation (5.6) cannot directly be determined because Equations (5.5) and (5.6) contain two unknowns. To measure out-of-plane (w) and in-plane (u, v) deformation components, a special adjustment (set-up) is required which will be described in the following part.

5.2.5 Measurement of Out-of-Plane and In-Plane Deformation

Out-of-plane deformation w can easily be reached by adjusting the illumination so that it is in the same direction as the observation direction. This causes the illumination angle α to become zero, thus eliminating the in-plane term $u(x, y)$ or $v(x, y)$ in Equations 5.5 and 5.6, and enabling the out-of-plane deformation $w(x, y)$ to be measured:

$$\Delta(x,y) = \frac{4\pi}{\lambda} w(x,y) \tag{5.7}$$

Figure 5.2c is obtained by an out-of-plane set-up. Each fringe depicts the same out-of-plane deformation with the maximum being at the centre (high fringe order) and the minimum around the outer edges (zero fringe order) for a square plate clamped all around and loaded centrally.

The adjustable range for the illumination angle α is from 0° to 90°. The term $(1+\cos\alpha)$ in Equations (5.5) and (5.6) can never be zero between this range of α. This indicates that in-plane deformation cannot be measured by the set-up shown in Figure 5.1. In order to measure the in-plane deformations, a dual-beam illumination method has been introduced [11]. Figure 5.3 shows the optical layout of ESPI, i.e. the well-known dual-beam illumination system, for in-plane deformation measurement. A laser beam is split into two beams to illuminate an object with a same illuminating angle $+\alpha$ to $-\alpha$ from left and right or from top to bottom. The light reflected from the object surface recombines at the position of a CCD camera, which results in a speckle interferogram.

Unlike the set-up shown in Figure 5.1, the in-plane set-up uses two beams to illuminate the object surface. The change of phase difference in the dual-beam illumination set-up due to a loading becomes:

$$\begin{aligned}\Delta(x,y) &= [\theta_1'(x,y) - \theta_2'(x,y)] - [\theta_1(x,y) - \theta_2(x,y)] \\ &= [\theta_1'(x,y) - \theta_1(x,y)] - [\theta_2'(x,y) - \theta_2(x,y)] \\ &= \Delta_1(x,y) - \Delta_2(x,y)\end{aligned} \tag{5.8}$$

where $\Delta_1(x,y)$ and $\Delta_2(x,y)$ are the phase changes of illumination beams 1 and 2 due to the loading, respectively. If the illumination angle α lies in the xoz-plane, then $\Delta_1(x,y)$ has a positive value of α and $\Delta_2(x,y)$ has a negative value of α, and Equation (5.8) can now be rewritten as:

$$\begin{aligned}\Delta(x,y) &= \Delta_1(x,y) - \Delta_2(x,y) \\ &= \frac{2\pi}{\lambda}\left\{\sin(\alpha)u(x,y) + \left[1+\cos(\alpha)\right]w(x,y)\right\} \\ &\quad - \frac{2\pi}{\lambda}\left\{\sin(-\alpha)u(x,y) + \left[1+\cos(-\alpha)\right]w(x,y)\right\} \\ &= \frac{4\pi}{\lambda}\sin\alpha_{\text{xoz}}u(x,y)\end{aligned} \tag{5.9}$$

Equation (5.9) shows that the change of phase difference in the dual-beam illumination set-up is only related to in-plane deformation, which indicates that the dual-beam illumination set-up is purely sensitive to in-plane deformation. Should the illumination angle α lies in the yoz-plane, Equation (5.9) becomes:

$$\Delta(x,y) = \frac{4\pi}{\lambda}\sin\alpha_{\text{yoz}}v(x,y) \tag{5.10}$$

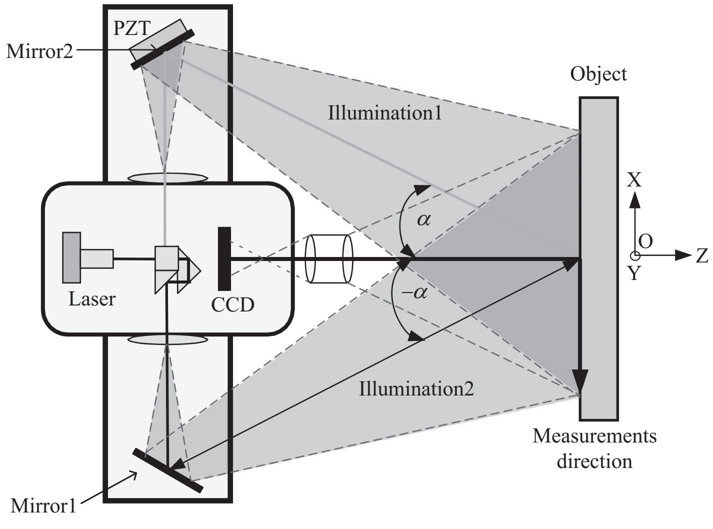


FIGURE 5.3 A schematic layout of ESPI set-up for in-plane displacement measurement.

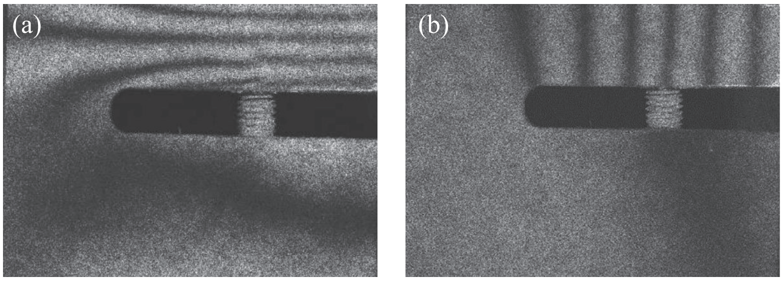


FIGURE 5.4 In-plane deformations *u*(*x,y*) (a) and *v*(*x,y*) (b) using dual-beam illumination for a plate with a notch loaded by a screw at the right end to expand the notch.

Figure 5.4 shows in-plane deformations $u(x, y)$ and $v(x, y)$ for an object with a horizontal notch loaded by a screw at the right side to expand the notch.

5.3 Evaluation of the Interference Phase in ESPI

The relationship between the phase change and deformation components has been described in Sections 5.2.4 and 5.2.5. The key point for measuring deformations is how to get the value of the phase change $\Delta(x, y)$at each pixel point, i.e. how to get the phase map of a fringe pattern. As described in Section 5.2.3, the phase change $\Delta(x, y)$ is equal to $N \times 2\pi$ at the locations of dark fringes, where $N=0, \pm 1, \pm 2, \pm 3$, etc, is the fringe order. Obviously, the smallest phase change $\Delta(x, y)$ between two adjacent dark fringes which can be measured is 2π. Therefore, the smallest out-of-plane and in-plane deformations which can be determined is $\lambda/2$ from Equation (5.7) and $\lambda/(2\sin\alpha)$ from Equation (5.10), respectively.

In order to reach high measurement sensitivity, i.e. to measure smaller deformations *u, v*, and *w*, a method to measure phase change within 2π is needed, and this is why the phase-shift techniques have been applied to ESPI [5,12–16]. Phase-shift techniques can be divided into two categories: temporal phase-shift technique, which has the higher fringe pattern quality, and spatial phase-shift technique, which has the higher dynamic range.

5.3.1 Temporal Phase-Shift Technique

In recent years, a variety of temporal phase-shift algorithms have been developed, including the 3+3, 3+1, 4+4 and 4+1 algorithms.

In order to measure the phase change $\Delta(x, y)$, the phase difference φ in the intensity equation should be measured first. As we know, a digital camera can record only intensity. With an 8-bit hardware resolution, the intensity values in the digital speckle interferometry lie between 0 and 255, where 0 is completely dark and 255 is completely white, with a total of 256 grey levels. In the intensity Equation (5.1), the only known term is the intensity I and other three terms, A, B, and the phase difference φ before a loading (or the phase difference after a loading φ'), are unknown. The temporal phase-shift measures phase φ (or φ') by recording three or more images in a time series so that three or more equations in a time series are obtained. In order to obtain three or more linearly independent equations, a constant and known phase shift is introduced between every two images. This phase shift is fulfilled by a piezoelectric transducer (PZT) driven mirror, and the driven mirror can be moved with precise displacement to generate a known phase shift which will be described in details later on.

5.3.1.1 4+4 Algorithm

Though 3 images (with three equations) can solve the phase φ, the 4 equations' phase-shift algorithm is the most widely used in the temporal phase-shift method [17]. This algorithm is fast and simple. Usually, it is called the 4+4 algorithm, i.e. 4 equations to solve φ (phase-before loading) and the other 4 to solve φ' (phase-after loading). A total of eight images are taken to calculate the phase change Δ. Before loading, four intensity images with a phase-shift increment of π/2 are captured. According to Equation (5.1), four intensity equations can be expressed as follows:

$$\begin{aligned} I_1 &= A + B\cos\varphi \\ I_2 &= A + B\cos(\varphi + \pi/2) \\ I_3 &= A + B\cos(\varphi + \pi) \\ I_4 &= A + B\cos(\varphi + 3\pi/2) \end{aligned} \tag{5.11}$$

The phase-before loading can be calculated as

$$\varphi = \arctan\frac{I_4 - I_2}{I_1 - I_3} \tag{5.12}$$

Similarity, the same process can be performed to calculate the phase-after loading:

$$\varphi' = \arctan\frac{I_4' - I_2'}{I_1' - I_3'} \tag{5.13}$$

where I_1', I_2', I_3' and I_4' are the intensity of 4 images after loading. Then, the phase change Δ can be determined by subtracting φ from φ':

$$\Delta = \begin{cases} \varphi' - \varphi, & \text{if } \varphi' > \varphi \\ \varphi' - \varphi + 2\pi, & \text{else} \end{cases} \tag{5.14}$$

Then, by substituting the phase change Δ in Equations (5.7), (5.9) or (5.10), the out-of-plane deformation *w* and the in-plane deformations *u* and *v* can be evaluated accurately.

5.3.1.2 4+1 Algorithm

Since the 4+4 temporal phase-shift algorithm requires four images at each loading condition, it is not well suited for dynamic tests (except for harmonic exciting vibration measurements with stroboscopic illumination [18]). The 4+1 fast temporal phase-shift algorithm was proposed to fulfil the dynamic measurement capability [19]. Four images are taken using the same procedure as described in the last section before loading. During loading, the camera only records a single image.

The images captured using the four-step phase-shift method before loading can be represented by Equation (5.11), and the recorded image during loading can be expressed as $I' = A + B\cos(\varphi + \Delta)$. Subtracting each intensity equation of (5.11) from the loading intensity equation I', then taking the square of the difference results in the following equation group:

$$\begin{aligned}
\overline{I}_1^{\,2} &= \left(I' - I_1\right)^2 = \left[2B\sin\left(\varphi + \Delta/2\right)\sin\Delta/2\right]^2 \\
\overline{I}_2^{\,2} &= \left(I' - I_2\right)^2 = \left[-2B\cos\left(\varphi + \Delta/2 - \pi/4\right)\sin\left(\pi/4 - \Delta/2\right)\right]^2 \\
\overline{I}_3^{\,2} &= \left(I' - I_3\right)^2 = \left[-2B\cos\left(\varphi + \Delta/2\right)\cos\Delta/2\right]^2 \\
\overline{I}_4^{\,2} &= \left(I' - I_4\right)^2 = \left[2B\sin\left(\varphi + \Delta/2 - \pi/4\right)\cos\left(\pi/4 - \Delta/2\right)\right]^2
\end{aligned} \tag{5.15}$$

Though $\Delta/2$ is a low-frequency component (induced by a loading), the phase difference φ between object and reference beams is a high-frequency component due to a surface roughness as described in Section 5.2.3. Additionally, the terms $\varphi + \Delta/2$, $\left(\varphi + \Delta/2 + \pi/4\right)$ and $\left(\varphi + \Delta/2 - \pi/4\right)$ are high-frequency components. For the high-frequency terms, a detector captures an average value of the term within each speckle [20]. With this assumption, we have the following result:

$$\left\langle \sin^2\left(\frac{\Delta}{2} + \varphi\right)\right\rangle \approx \frac{1}{2n\pi}\int_0^{2n\pi} \sin^2\left(\frac{\Delta}{2} + \varphi\right) d\left(\frac{\Delta}{2} + \varphi\right) = \frac{1}{2} \tag{5.16}$$

Similarly, $\left[\sin\left(\varphi + \Delta/2 - \pi/4\right)\right]^{2}{}^{2}$, $\left[\cos\varphi + \Delta/2\right]^2$, and $\left[\cos\left(\varphi + \Delta/2 - \pi/4\right)\right]^2$ are also equal to 1/2. Therefore, Equations 5.15 can be simplified to:

$$\begin{aligned}
\overline{I}_1^{\,2} &= B^2\left[1 - \cos\Delta\right] \\
\overline{I}_2^{\,2} &= B^2\left[1 - \cos\left(\Delta + \pi/2\right)\right] \\
\overline{I}_3^{\,2} &= B^2\left[1 - \cos(\Delta + \pi)\right] \\
\overline{I}_4^{\,2} &= B^2\left[1 - \cos\left(\Delta + 3\pi/2\right)\right]
\end{aligned} \tag{5.17}$$

The phase change Δ can be calculated by solving the equation group (5.17):

$$\Delta = \arctan\frac{\overline{I}_4^{\,2} - \overline{I}_2^{\,2}}{\overline{I}_1^{\,2} - \overline{I}_3^{\,2}}. \tag{5.18}$$

The 4+1 phase-shift algorithm provides the capability to realize dynamic measurements using the temporal phase-shift technique. However, the phase-map quality is not as good as the 4+4 method due to the assumption that is shown in Equation (5.16). Therefore, the 4+1 algorithm is not widely used in high-accuracy quantitative measurement but can be utilized in some qualitative NDT applications with dynamic loading Figure 5.5 shows a comparison of two phase maps taken from the same set-up, but using different algorithms. Both were smoothed by the same times. The object is a square plate clamped all around and loaded centrally. It is obvious that 4+4 algorithm provides a much better phase map quality than the 4+1 algorithm.

5.3.1.3 Other Temporal Phase-Shift Algorithms

Besides these two algorithms mentioned in the previous sections, there are other different temporal phase-shift algorithms [21–23]. Some of them will be briefly introduced in this section.

The first one is the 3+3 temporal phase-shift algorithm. As described in the previous section, three unknowns exist in the intensity equation; thus, at least three equations are required to solve the phase φ. In the 3+3 temporal phase-shift algorithm, three images with an introduced phase shift of $3\pi/2$ between the images are captured at each loading condition. The 3+3 temporal phase-shift algorithm provides similar phase map quality as the 4+4 temporal phase-shift algorithm. The algorithm for calculating phase is not as simple as the 4+4 algorithm. Like the 4+1 fast temporal phase-shift algorithm, a 3+1 fast temporal phase-shift algorithm also exists to achieve NDT applications with dynamic measurement.

Other than the 4+4 and 3+3 standard temporal phase-shift algorithms and the 4+1 and 3+1 fast temporal phase-shift algorithms, 5+5 algorithms and algorithms of even more images also exist. Usually, the more images, the more accurate. However, in speckle interferometry, the relative phase change,

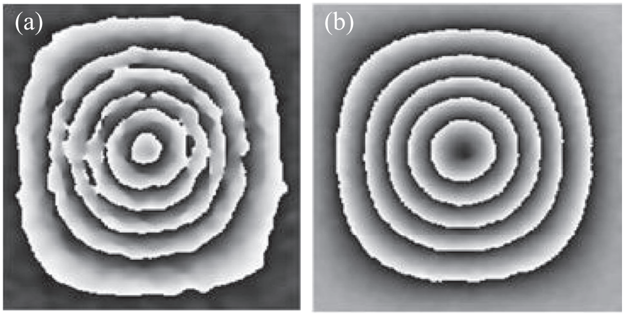


FIGURE 5.5 A comparison of phase maps using 4+1 and 4+4 algorithms (after two-time phase smooths with a 5 x 5 window). The test object is a square plate clamped all around and loaded centrally: (a) 4+1 method; (b) 4+4 method.

which is obtained by the subtraction of two calculated phase distributions before and after loading, is of interest. Some systematic errors will be eliminated by the subtraction function, and the 4+4 algorithm is still the most popular method for temporal phase-shift technique.

5.3.1.4 Generation of a Phase-Shift

One important task in the temporal phase-shift technique is how to generate a phase shift. Usually, it is achieved by moving a mirror driven by a PZT [24]. An optical path change of a wavelength corresponds to a phase change of 2π. Based on this relationship, a phase shift can be generated by controlling the movement of the PZT mirror.

Figure 5.6 shows three typical set-ups for generating phase-shift. Set-ups (a), (b) and (c) are suitable set-ups if their optical propagation directions are 180°, 90°, and 0° (no direction change), respectively. It should be emphasized that Figure 5.6b will produce an optical beam lateral offset while moving the PZT mirror, which will affect the interference result. 5.6c uses a hollow roof prism mirror, resulting in a better set-up that can be placed on any light path without changing the direction and without generating a lateral offset while moving the mirror. Besides these set-ups shown, a PZT ring can also generate a phase-shift if light propagates through a fibre (fibres). In this case, the fibre is wound around the PZT ring. When the diameter of the PZT ring changes slightly, a phase change will be generated because of fibre expansion. Calibration is necessary for such a set-up [25–26].

5.3.2 Spatial Phase-Shift Technique

The temporal phase-shift method extracts phase from multiple images recorded at different times with a known phase shift, while the spatial phase-shift method extracts phase from a single image. Therefore, the spatial phase-shift method is more suitable for dynamic measurement, especially for transient loading conditions if a high-speed camera will be utilized.

5.3.2.1 Spatial Phase-Shift Method Based on Multi-Pixel Calculation

In order to calculate phase distribution with a single image, a spatial phase-shift method based on multi-pixel calculation will be introduced in this part.

Instead of recording multiple images as is done in the temporal phase-shift method, the principle of the spatial phase-shift method based on multi-pixel calculation is to generate a phase difference among three neighbouring pixels by adjusting the reference beam. Figure 5.7 is a schematic of the idea. The red beam is the object beam and the blue beam is the reference beam in ESPI. By setting the reference beam at an appropriate angle α, an optical path difference (OPD) between two adjacent pixels is generated:

$$OPD = S \cdot \sin\alpha \tag{5.19}$$

where S is the pixel size. If the laser wavelength is λ, a phase-shift δ then is obtained by

$$\delta = \left(\frac{OPD}{\lambda}\right)2\pi = \left(S \cdot \frac{\sin\alpha}{\lambda}\right)2\pi \tag{5.20}$$

Using the three-equation algorithm as an example, by adjusting the reference beam with a suitable angle α, a phase difference δ between every two pixels, based on Equation 5.20, is obtained and three intensity equations among three neighbouring pixels are generated:

$$\begin{aligned} I_1 &= A + B\cos\varphi \\ I_2 &= A + B\cos(\varphi + \delta) \\ I_3 &= A + B\cos(\varphi + 2\delta) \end{aligned} \tag{5.21}$$

If the phase-shift δ is selected as $(2N\pi + 2\pi/3)$ by adjusting the angle α appropriately, the phase difference φ can be calculated by:

$$\varphi = \arctan\frac{\sqrt{3}(I_1 - I_3)}{2I_2 - I_1 - I_3} \tag{5.22}$$

N in the equation of δ is an integer number 0, 1, 2 3, etc., that is selected depending on easy adjustment of the reference beam angle α. Repeating the same procedure after the object is loaded, the phase difference φ' can be calculated and the phase change Δ related to the deformation can be determined as presented in Equation 5.16.

The values A, B, and φ (see Equation 5.21) must be the same among the three neighbouring pixels to use the spatial phase-shift method based on multi-pixel calculation. In order to meet these conditions, the aperture size of the camera should be adjusted appropriately so that each speckle covers at least three pixels. The three neighbouring pixels will have the same A, B and φ because they share the same speckle [27].

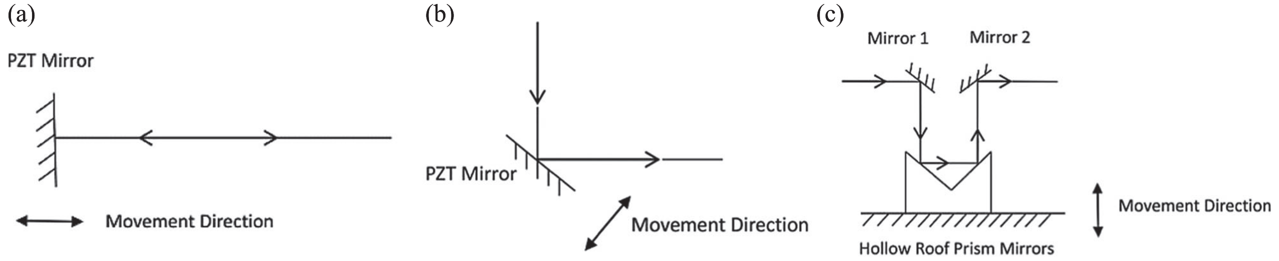


FIGURE 5.6 Different methods for generating phase-shift: (a) a PZT mirror changing 180° propagation direction, (b) a PZT mirror changing 90° propagation direction, (c) hollow roof prism mirrors without changing propagation direction.

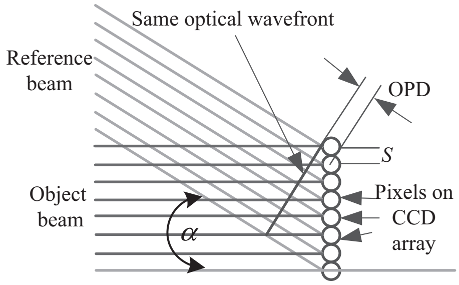


FIGURE 5.7 Schematic of multiple-channel spatial phase-shift among pixels.

The method described above is also called the spatial phase-shift method based on multi-channels. Instead of multiple pixels, some researchers used multiple cameras to simultaneously record multiple images, and a phase shift is introduced between each camera by using several polarization components [28,29]. The set-up for the multi-camera system, especially the generation of the phase-shift between those cameras, is hard to achieve in practice, and the relative movement between cameras caused by vibration or temperature change will generate large errors in this system. The cost of this approach is also high, so there have been few real applications for this approach. Another multi-channel system approach uses a single CCD, but the CCD is divided into four regions [30]. A speckle pattern is imaged by a special imaging system at the four regions of the CCD simultaneously, and a phase-shift is introduced between each region. Though this method can be used for dynamic deformation measurement, the resolution is low because only a quarter of the CCD is used. Additionally, the imaging system and the phase-shifting system are complex, which makes the system difficult to achieve for real applications.

5.3.2.2 Spatial Phase-Shift Method Based on Carrier-Frequency and Fourier Transformation

The other important spatial phase-shift method in ESPI is based on carrier frequency and Fourier transformation [31,32]. The principle of the method is illustrated in Figure 5.8. A laser is divided by a beam splitter into two beams, namely, an object beam and reference beam. The object beam illuminates an object after it is expanded. Then, the scattered light is collected by a positive lens which forms an image on a CCD sensor. An aperture is placed in front of the positive lens and serves to limit the spatial frequencies of the object wave. Being carried by a single-mode fibre, the reference beam strikes the CCD sensor and interferes with the object beam. As a result, a speckle pattern is formed and captured by the CCD. Let $U(x, y)$ be the object wave and $R(x, y)$ be the reference wave. The intensity distribution of the speckle pattern is then given by:

$$\begin{aligned} I(x,y) &= \left[R(x,y)+U(x,y)\right]\left[R^*(x,y)+U^*(x,y)\right] \\ &= \left|R(x,y)\right|^2+\left|U(x,y)\right|^2+R(x,y)U^*(x,y)+R^*(x,y)U(x,y) \end{aligned} \tag{5.23}$$

where * denotes the complex conjugate and (x, y) is the two-dimensional space coordinates on the CCD plane. Different spatial frequencies of the speckle patterns are formed when various inclination angles between the object and reference beams are introduced. The angle α shown in Figure 5.8 is the inclination angle between the object and reference beams. The carrier frequency introduced by the angle can be expressed by:

$$f_c = \frac{\sin\alpha}{\lambda} \tag{5.24}$$

where f_c is the carrier frequency of speckle pattern and λ is the laser wavelength.

A Fourier transform and inverse Fourier transform are used to extract the interferometric phases from the speckle pattern. When applying the Fourier transform to Equation (5.23), it becomes,

$$I(f_x,f_y) = \mathrm{FT}[\mathrm{I}(x,y)] = A(f_x,f_y) + B(f_x,f_y) + C(f_x,f_y) \tag{5.25}$$

where $I(f_x, f_y)$ is the distribution on the frequency domain, (f_x, f_y) is the coordinate in the frequency domain, the symbol FT denotes the processing of Fourier transform, and

$$\begin{aligned} A(f_x,f_y) &= \mathrm{FT}\left[\left|R(x,y)\right|^2+\left|U(x,y)\right|^2\right] \\ B(f_x,f_y) &= \mathrm{FT}\left[R(x,y)U^*(x,y)\right] \\ C(f_x,f_y) &= \mathrm{FT}\left[R^*(x,y)U(x,y)\right] \end{aligned} \tag{5.26}$$

Equation (5.25) shows that the spatial frequency distribution is separated into three components: $A(f_x, f_y)$, $B(f_x, f_y)$ and $C(f_x, f_y)$, where $B(f_x, f_y)$ and $C(f_x, f_y)$ contain the complex amplitude of

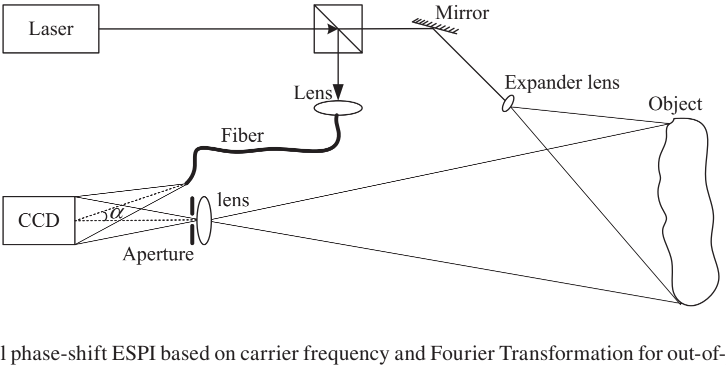


FIGURE 5.8 Principle of spatial phase-shift ESPI based on carrier frequency and Fourier Transformation for out-of-plane deformation measurement.

the object beam and $A(f_x, f_y)$ only contains the background light. To demonstrate the concepts described in this section, an example is presented in Figure 5.9. The object to be tested is a square plate. Figure 5.9a shows a speckle pattern of a square plate after introducing the carrier frequency captured by the set-up shown in Figure 5.8. The carrier fringes can be observed from a zoom-enlarged image. This image corresponds to the descriptions of Equations 5.23. Figure 5.9b shows the spectrum after applying an FT on Figure 5.9a. There are three separate parts, labelled A, B and C, marked by three red circles in the picture. Part A is a low-frequency component containing background information that is useless for the measurement. A filtering window can be utilized to extract the phase either from component B or C.

The complex amplitude of the object beam can be obtained by making an inverse Fourier transform of $C(f_x, f_y)$ or $B(f_x, f_y)$:

$$R^*(x,y)U(x,y) = \mathrm{IFT}\left[C\left(f_x, f_y\right)\right] \tag{5.27}$$

and then the phase can be calculated by:

$$\varphi(x,y) = \frac{\mathrm{Im}\left[R^*(x,y)U(x,y)\right]}{\mathrm{Re}\left[R^*(x,y)U(x,y)\right]} \tag{5.28}$$

Repeating the same procedure, the phase difference φ' after loading can be calculated, and the phase change Δ can be determined. The phase change Δ in the set-up shown in Figure 5.9 is related to the out-of-plane deformation if the angle between the illumination and observation directions is equal or close to zero. Thus, the set-up is also called out-of-plane spatial phase-shift ESPI system.

For measuring pure in-plane deformation, the well-known dual beam illumination as shown in Figure 5.3 should be utilized. However, the spatial phase-shift method based on carrier frequency cannot be introduced for a standard dual-beam system because one of two scattered light beams needs to be adjusted to introduce a carrier frequency, and the scattered light from two beams is mixed together and cannot be separated.

A novel spatial phase-shift dual-beam illumination ESPI system, which is sensitive to pure in-plane deformation, has been reported recently [33]. Figure 5.10 presents the principle of this new idea. A low-coherence laser is split into two beams by a variable beam splitter (S): one beam illuminates the object in the direction of the vector k_1 after being expanded by the expander BE1, while the other beam illuminates the object after the mirror M1 and the expander BE2 in the direction of the vector k_2. The observation direction of the CCD is in the direction of the vector k_3. The angles of the observation direction and two illumination directions are α and $-\alpha$, respectively. The scattered light from the object surface from k_1 to k_2 illumination directions is referred as (beam-k_1) and (beam-k_2), respectively. If the BS1 is a 50:50 beam splitter, 50% on the total intensity of the scattered light from beam-k_1 and beam-k_2 reaches the aperture A1, marked as (beam-k_1)$_{A1}$ and (beam-k_2)$_{A1}$ and the other 50% reaches the aperture A2, marked as (beam-k_1)$_{A2}$ and (beam-k_2)$_{A2}$. The optical path from the beam-k_2 is longer than the beam-k_1. The difference is equal to the distance between S and M1 in the figure, labelled as S-M1. When selecting a laser use, a low-coherence laser whose coherent length is shorter than S-M1 should be utilized. With this condition, the two beams (beam-k_1)$_{A1}$ and (beam-k_2)$_{A1}$ cannot interfere with each other before the aperture A1 because their OPD is larger than the coherent length of the laser. Similarly, the two beams (beam-k_1)$_{A2}$ and (beam-k_2)$_{A2}$ cannot interfere with each other before the aperture A2. However, if an additional optical path, equal to S-M1, is introduced before A2, then the (beam-k_1)$_{A2}$ in front of aperture A2 has the same optical path as the (beam-k_2)$_{A1}$ in front of aperture A1. They can interfere with each other when they reach the CCD. The mirrors 3–5 are used just for this purpose. Furthermore, for the two beams that are separated by A1 and A2, a carrier fringe can be easily introduced by an angle between the two apertures when they are mapped to the same optical axis. Therefore, a novel spatial phase-shift dual-beam ESPI system which is sensitive to pure in-plane deformation is achieved. Please note that the (beam-k_2)$_{A2}$ in the front of aperture A2 has an even larger optical path than the (beam-k_1)$_{A1}$ in the front of aperture A1, and they cannot interfere with each other when they reach the CCD. The phase extraction process is the same as that described for the out-of-plane spatial phase-shift ESPI system.

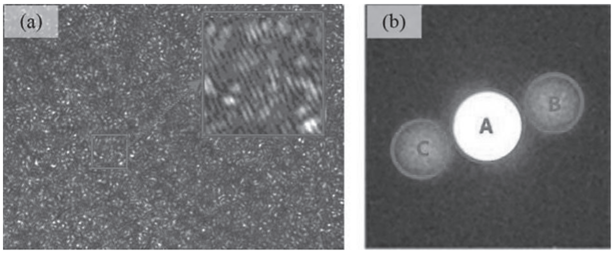


FIGURE 5.9 (a) Speckle pattern of spatial phase-shift ESPI, the region within the rectangular box is magnified to show the carrier fringes; (b) spectrum obtained after Fourier transform.

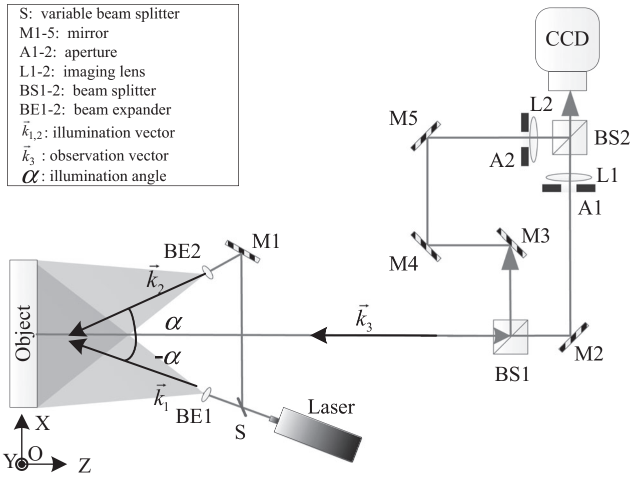


FIGURE 5.10 Schematic of spatial phase-shift dual-beam illumination for in-plane measurement.

If the illumination angles α and $-\alpha$ lie in the XOZ plane, then in-plane deformation in the x-direction u can be measured. If they lie in the YOZ plane, in-plane deformation in the y-direction v can be measured.

As described in this section, the spatial phase-shift methods can be divided into two main categories: multi-pixel/multi-channel method and carrier frequency with Fourier Transform method. The multi-pixel approach can determine phase distribution from a single speckle pattern, making it suitable for dynamic measurement. However, three or more pixels are needed to determine the phase distribution of one point, greatly reducing the spatial resolution. Additionally, this approach can only conduct measurements in one dimension, such as the out-of-plane direction. A carrier frequency method provides higher spatial resolution. It is capable of measuring multiple deformation components by recording multiple interferograms from different speckle interferometers in a single image [34]. The interferograms can then be separated in the frequency domain by using different carrier frequencies, and 3D deformations can be extracted from the interferograms, which will be described in detail in the application part of the next section.

5.4 Applications of ESPI

The main applications of ESPI are for deformation measurements, including out-of-plane, in-plane and three-dimensional deformation measurements, for vibration measurement, and for non-destructive testing. With the introduction of phase-shift techniques, evaluation of ESPI fringes becomes automatic, quantitative and more accurate, which opens numerous possibilities for strain measurement, vibration analysis and non-destructive evaluation. In this part, applications of ESPI will be described and demonstrated using examples.

5.4.1 Applications of Temporal Phase-Shift ESPI

Temporal phase-shift ESPI is mainly applied for deformation and strain measurement under static loadings and non-destructive testing. It can also be utilized for vibration analysis through stroboscopic illumination, which will be described in a separate section in 5.4.3.

5.4.1.1 Out-of-Plane and In-Plane Deformation Measurement and Non-destructive-Test

Figure 5.11 shows the measurement of out-of-plane deformation by using the set-up presented in Figure 5.1 under a zero-elimination angle. The temporal phase-shift algorithm used is a 4+4 method. The object to be tested is a square plate clamped all around and loaded centrally. Figure 5.11a and b shows 4 intensity images before and after loading with a phase shift of 90° between every two successive images and the calculated phase distribution based on Equations 5.12 and 5.13, respectively. Figure 5.13c–f presents the original phase map obtained by subtracting φ' from φ based on Equation 5.14, the filtered phase map, the unwrapped phase map and the quantitative 3D evaluation of the out-of-plane deformation *w*, respectively. The details regarding the phase-map smoothing and unwrapping can be found from references [35,36].

As shown in Equation 5.7, ESPI measures the deformation *w* by determining the phase change Δ. Therefore, the sensitivity for measuring deformation *w* depends on the measurement sensitivity for phase determination. For the fringe based on intensity subtraction shown in Figure 5.2, the smallest phase that can be determined is 2π. However, the sensitivity for phase measurement can be much smaller using the phase-shift technique. The phase measurement sensitivity ranges from 2π/10 to 2π/30, depending on a number of factors including speckle noise, PZT set-up and the moving accuracy, the hardware resolution, different algorithms and smoothing methods adopted [37,38]. For an averaged phase-measuring sensitivity of 2π/20, the sensitivity for determining an out-of-plane deformation *w* can reach as small as 15 nm if the laser wavelength

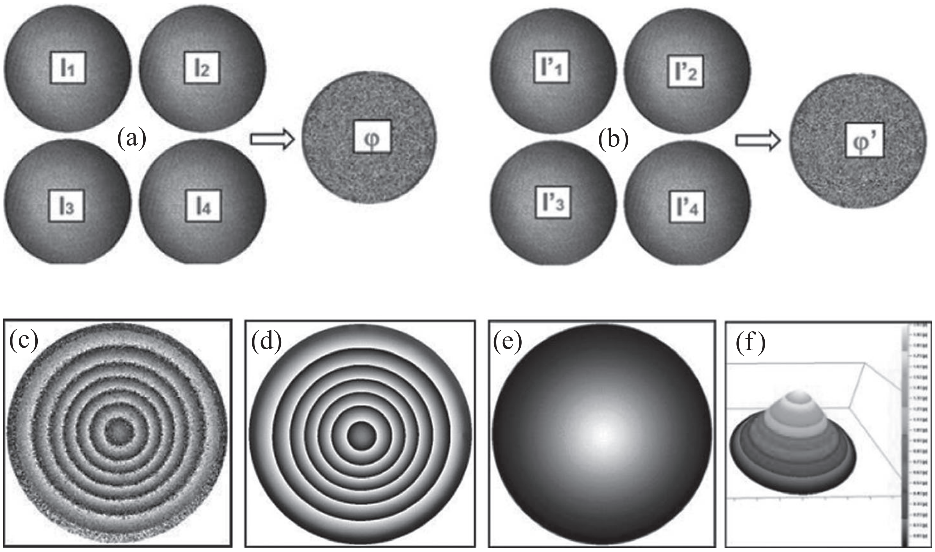


FIGURE 5.11 (a) 4 intensity images and the calculated phase distribution φ before loading; (b) other 4 intensity images and the calculated phase distribution φ' after loading; (c) the original phase map of the phase change Δ; (d) the filtered phase map, (e) the unwrapped phase map and (f) the quantitative 3D evaluation of the out-of-plane deformation *w*.

used is 600 nm. For NDT applications, this means that smaller defects can be found.

A comparison between ESPI fringe for NDT applications determined with and without applying the phase-shift technique is shown in Figure 5.12. The object being tested is a sealed composite container loaded by an internal pressure. Two de-laminations were found in the ESPI fringe obtained by the intensity subtraction technique, whereas three de-laminations were clearly demonstrated in the phase map of ESPI fringe using the phase-shift technique. It is a fact that a visible delamination should generate at least one fringe (or a 2π phase change) in the intensity fringe due to a loading, whereas a phase change lower than 2π can be measured in the phase-shift technique and a smaller delamination can subsequently be found.

An application for in-plane deformation measurement using two dual-beam illumination systems (one measuring *u* and the other measuring *v*) is shown in Figure 5.13. The shear bands in aluminium during a tensile testing have clearly been demonstrated. The top and bottom left pictures show the phase map of deformation *u & v*, respectively. The middle and right images are the corresponding wrapped deformation and strain images. More applications for in-plane deformation and strain measurement can be found in reference [39–40].

5.4.1.2 Measurement of Three-Dimensional Deformations and Strain

An approach to measure three-dimensional (3D) deformation is to take one out-of-plane set-up to measure out-of-plane deformation *w* and two dual-beam illumination ESPI set-ups, one in the xoz plane and the other in the yoz plane, to measure in-plane deformations *u* and *v*. By utilizing three optical switches, 3D deformation can be measured under a single load [34,37]. Strains are determined by numerically differentiating the deformation data. Like most experimental techniques, ESPI measures only surface information, so the deformations *u, v* and *w* are functions of *x* and *y* only. Therefore, only in-plane strains $\varepsilon_x(\partial u/\partial x), \varepsilon_y(\partial u/\partial x)$ and shear strain $\gamma_{xy}(\partial u/\partial y + \partial v/\partial x)$ can be determined.

Figure 5.14 demonstrates the measurement of 3D-displacements and strains of a flat aluminium plate with a groove. As shown in Figure 5.14a, the plate is loaded on the bottom right side by driving in a screw, while three phase maps related to *u*(*x, y*), *v*(*x, y*) and *w*(*x, y*) are shown in 5.14b–d, respectively. The corresponding 3D deformations *u*(*x, y*), *v*(*x, y*) and *w*(*x, y*), obtained by multiplying the unwrapped phase maps by the system constants (cf. Equations 5.9 and 5.11), are presented in Figures 5.14e–g, respectively, while the in-plane strains $\varepsilon_x, \varepsilon_y$ and the shear strain γ_{xy}, determined through numerical differentiation of the deformation data, are demonstrated in Figures 5.14h–j.

It should be noted that the approach described above to measure 3D deformations is only suited for flat surfaces. Surface shape information is required to measure 3D deformations on a curved surface because coordinate systems on the object surface, defining the in-plane and out-of-plane components, vary with the contour as shown in Figure 5.15. Usually, the three combined interferometers' approach has a same coordinate system, known as the sensor coordinate system. For this

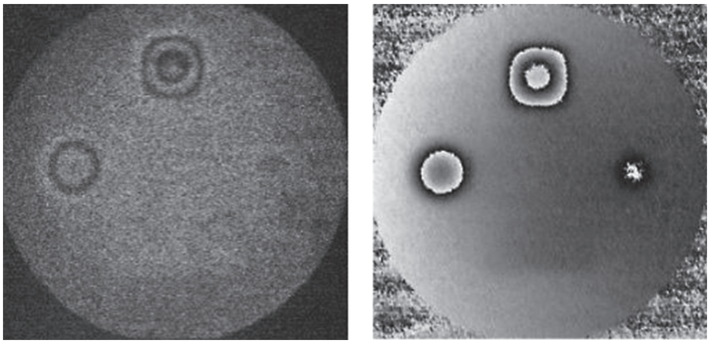


FIGURE 5.12 A comparison between ESPI fringe without (left) and with (right) using phase-shift technique by a same loading: (a) two de-laminations were found in the intensity fringes and (b) three delaminations were clearly demonstrated in the phase map of the fringes.

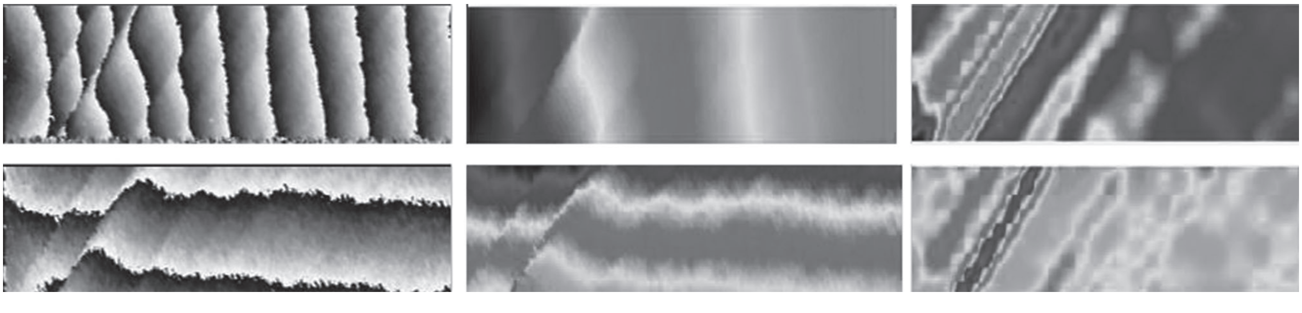


FIGURE 5.13 In-plane deformation measurement using two dual-beam illumination systems, phase maps *u* (top left) and *v* (bottom left), the corresponding deformations (middle) and the strains (right).

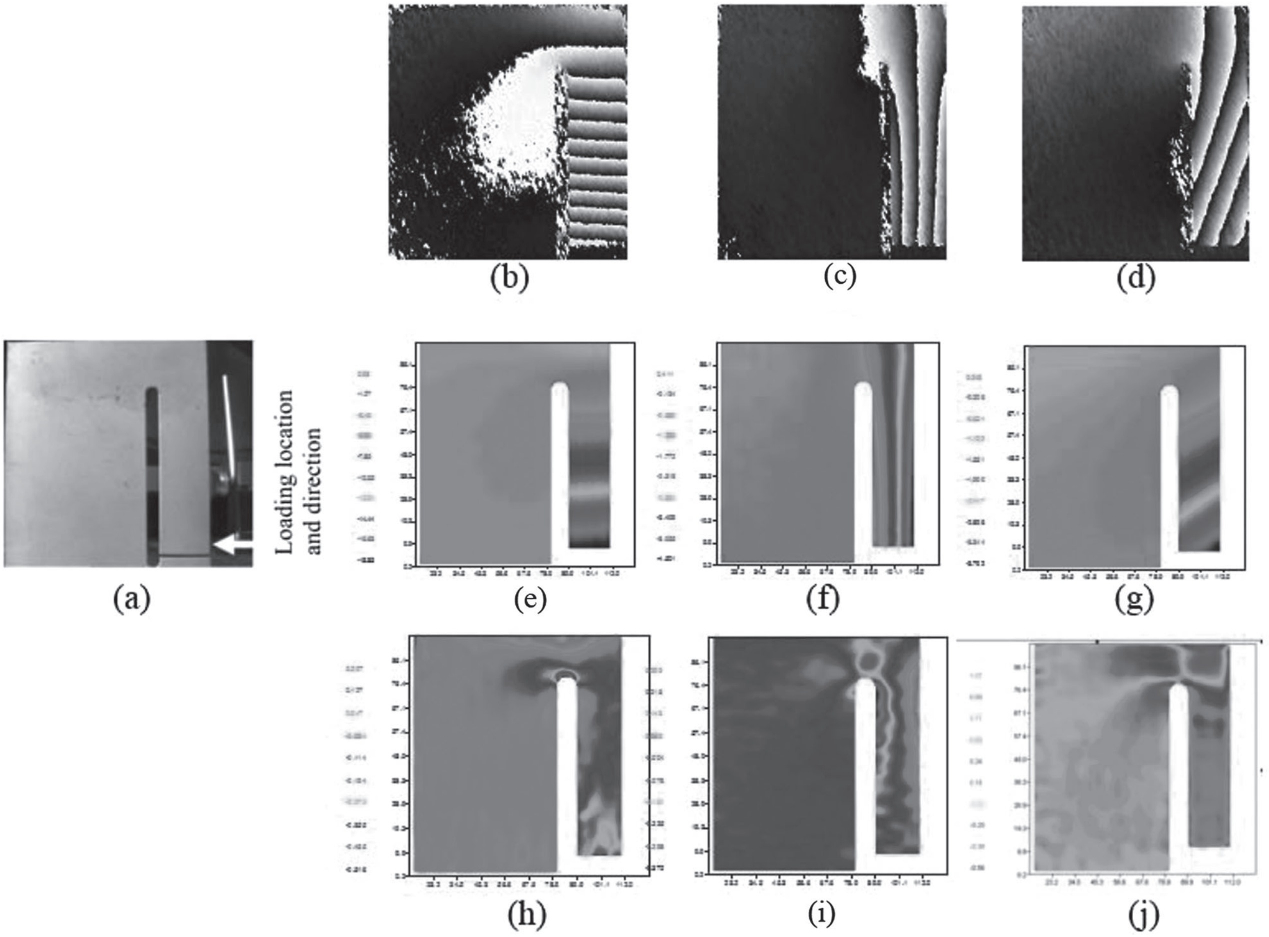


FIGURE 5.14 Measurements of 3D-displacements and strains of a flat aluminium plat with a groove loaded at the bottom of the right side (a) based on three interferometers combined approach. (b), (c) and (d): three phase maps in *x, y* and *z* directions, respectively; (e), (f) and (g): the corresponding three deformations *u, v* and *w*, respectively; (h), (i) and (j): in-plane strains $\varepsilon_x, \varepsilon_y$ and the shear strain γ_{xy}, respectively.

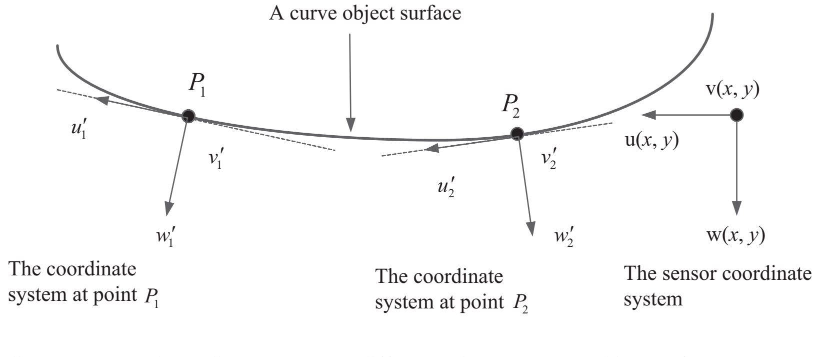

A curve object surface
P_1
P_2
v(x, y)
u(x, y)
w(x, y)
u_1'
v_1'
w_1'
u_2'
v_2'
w_2'
The coordinate system at point P_1
The coordinate system at point P_2
The sensor coordinate system

FIGURE 5.15 Sensor coordinate system and coordinate systems at different points on a curve object surface.

discussion, the sensor coordinate system is denoted as Oxyz and the corresponding deformations as *u*(*x, y*), *v*(*x, y*) and *w*(*x, y*). When the object under test has a curved surface, the coordinate systems on the object surface, denoted as Ox′y′z′ (usually the z' axis is normal to object surface), differ from the sensor coordinate system. The corresponding deformations on the object surface are denoted as *u'*(*x, y*), *v'*(*x, y*) and *w'*(*x, y*).

Obviously, the 3D-deformations on the object surface differ from the measured 3D deformation on the sensor coordinate system. The 3D-deformations *u'*, *v'* and *w'* on the object surface can, however, be determined from the measured 3D deformation *u*, *v* and *w* on the sensor coordinate system by applying the coordinate transformation equation [41]:

$$\begin{pmatrix} u' \\ v' \\ w' \end{pmatrix} = \begin{pmatrix} l_1,\ m_1,\ n_1 \\ l_2,\ m_2,\ n_2 \\ l_3,\ m_3,\ n_3 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} \tag{5.29}$$

TABLE 5.1

Direction Cosine Values between the Sensor and Object Coordinate Systems

	x	y	z
x'	l_1	m_1	n_1
y'	l_2	m_2	n_2
z'	l_3	m_3	n_3

where *l, m and n* are the directional cosine values, given in Table 5.1, between two coordinate systems:

An object-shape measurement is required to determine the values of these direction cosines. Different ESPI systems for 3D shape measurement have been developed [41–44]. The most popular ESPI system for contour measurement is the dual-beam illumination method. Instead of a loading between two images, an optical path change is introduced by moving a certain mirror; this causes the interference fringes to be related to the 3D shape. Once 3D shape information is obtained, the directional cosine values can be calculated [45]. The details for determining strains in combination with deformation and contour data can be found in reference [26].

Figure 5.16 shows an example of 3D deformation and strain measurement on a curved surface. The object being tested is a mouse femur that is positioned in a sample table of a custom-made piezoelectric mechanical loader. White powders were thinly sprayed onto the femoral surface during measurement to increase surface reflectance. A T-form aluminium rod driven by a PZT loader was used for loading (cf. the live image). The 3D-deformations and the 3D-shape measured are displayed in Figure 5.16a and b, respectively, and the strains $\varepsilon_x, \varepsilon_y$ and γ_{xy} as well as the principal strains ε_{p1} and ε_{p2} are shown in Figure 5.16c.

Other applications, such as absolute deformation measurement and colourful ESPI system for measurement of 3D deformation, can be found in reference [34,37,46–50].

5.4.2 Applications of Spatial Phase-Shift ESPI

Spatial phase-shift ESPI is mainly applied for deformation/strain measurement and NDT under non-static loadings, such as a continued, transient or a dynamic loading. It is also suited for simultaneously measuring 3D deformations under a single loading.

5.4.2.1 Out-of-Plane and In-Plane Deformation Measurement under a Continuous Loading

Figure 5.17a shows NDT results for a composite plate with three de-laminations using the out-of-plane spatial phase-shift ESPI set-up shown in Figure 5.8 under a continued vacuum loading. The object was tested in a vacuum chamber with a test speed of five images/second. All three de-laminations were found with increased load. The test was also conducted using the 4+1 temporal phase-shift ESPI system, as described in Section 5.3.1.2 using same loading speed. The results for this test are presented in Figure 5.17b. A comparison of the measurement results shows that the quality of the phase maps from spatial phase-shift ESPI is better than that for the 4+1 temporal phase-shift system. This results from the special assumption, as indicated in Equation 5.16, within the algorithm of the 4+1 method.

An in-plane deformation measurement in a groove area using the spatial phase-shift dual-beam ESPI set-up described

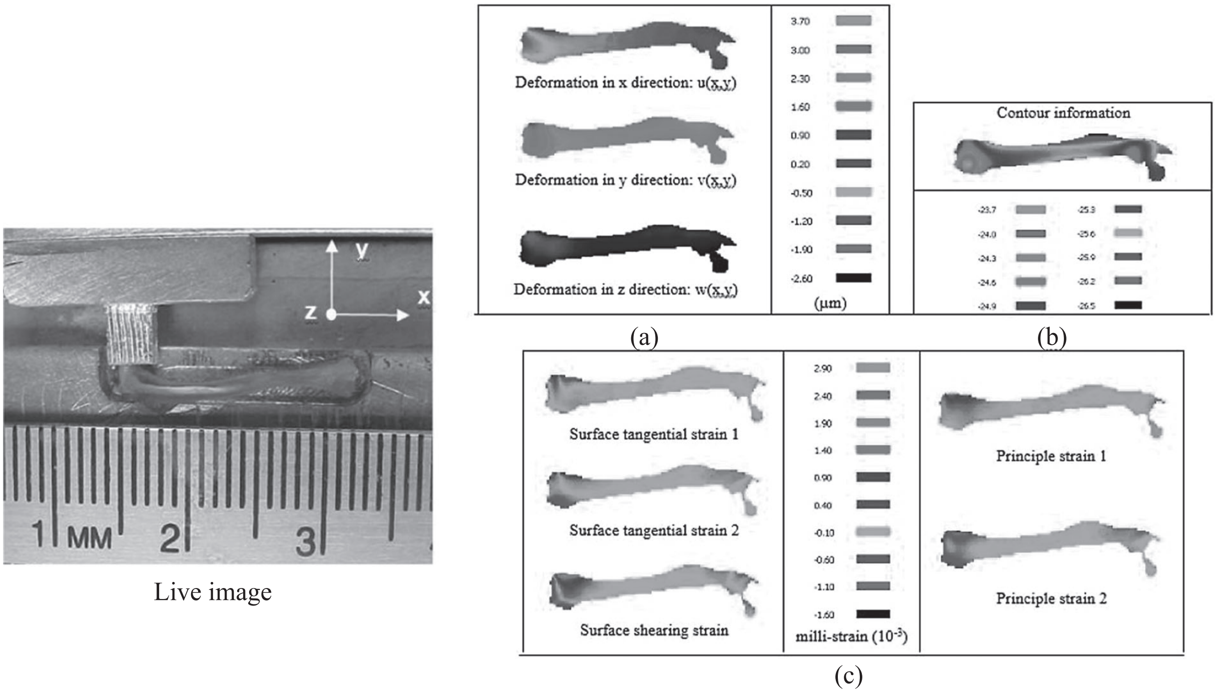


FIGURE 5.16 Deformation and strain measurement on a mouse femur (a curved surface): (a) the three components of deformation $u(x,y)$, $v(x,y)$ and $w(x,y)$ in the sensor coordinate system, (b) the 3D shape measurement and (c) strains $\varepsilon_x, \varepsilon_y$, and γ_{xy} as well as the principal strains ε_{p1} and ε_{p2}.

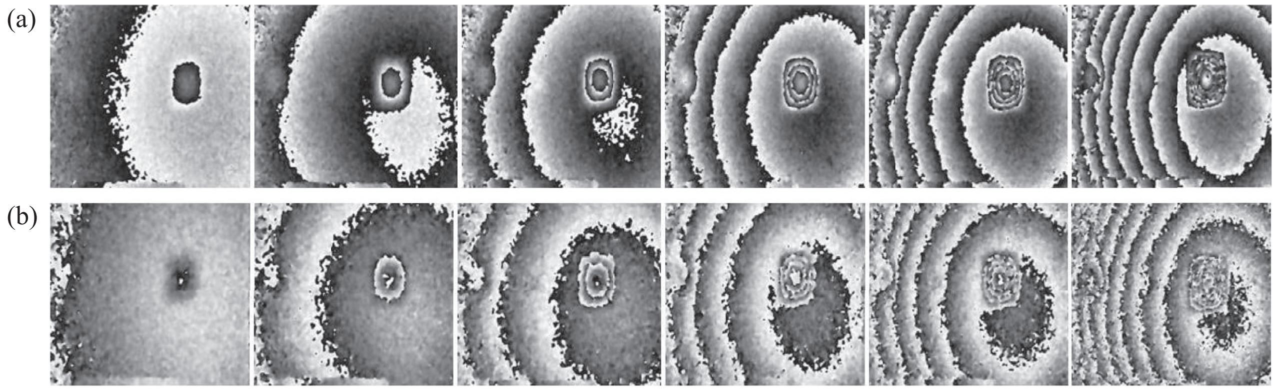


FIGURE 5.17 NDT for a composite plate with three de-laminations by a continued vacuum loading, the increasing direction of the loading is from left to right and a comparison between the spatial (a) and 4+1 temporal (b) methods.

in Figure 5.10 is presented in Figure 5.18. The dual-beam illuminations lie in the vertical plane. Thus, the in-plane deformation in the y-direction "v" was measured when the plate is loaded by driving a screw into its upside. The field view of the measured area is about 20 mm × 15mm, and the width of the groove is about 4 mm. The real-time fringe, the original phase map, the filtered phase map and in-plane deformation are demonstrated in Figures 5.18a–d, respectively. During the measurement, the loading was continuously increased, and speckle patterns were recorded at a rate of 15 frames per second. Selected deformations at different loading levels are shown in Figures 5.18e–h.

Other applications of spatial phase-shift ESPI for measuring displacement/deformation, vibration, etc., can be found in references [51–54].

5.4.2.2 Simultaneous Measurement of 3D-Deformations under a Single Loading

Simultaneous measurement of 3D-deformations under a single loading has been reported by using three interferometers with three different illumination (sensitivity) directions and three cameras, i.e. one camera for each interferometer. The three cameras are synchronized while taking the images. The spatial phase-shift method is used for phase measurement for each interferometer. Three interferometers with three illumination (sensitivity) directions generate three phase maps (i.e. three equations $\Delta_1(x, y)$, $\Delta_2(x, y)$ and $\Delta_3(x, y)$ with different sensitivity factors A, B and C as described in Equation 5.4). The three unknown deformation components *u*, *v* and *w* can thus be resolved from the three-phase change equation. The details regarding this study can be found in references [55–56].

In this part, a single camera with three spatial phase-shift ESPI set-ups for simultaneous measurement of 3D deformations is described. Figure 5.19a shows this experimental schematic diagram. Three 1D frequency carrier phase-shift ESPI set-ups are integrated together to measure the object from three different directions simultaneously. Three lasers with different wavelengths (A_1–A_3) are used to avoid interference between different measurement directions. The three reference beams are carried by optical fibres, each of them illuminating the CCD plane with a unique angle. In this case, three different

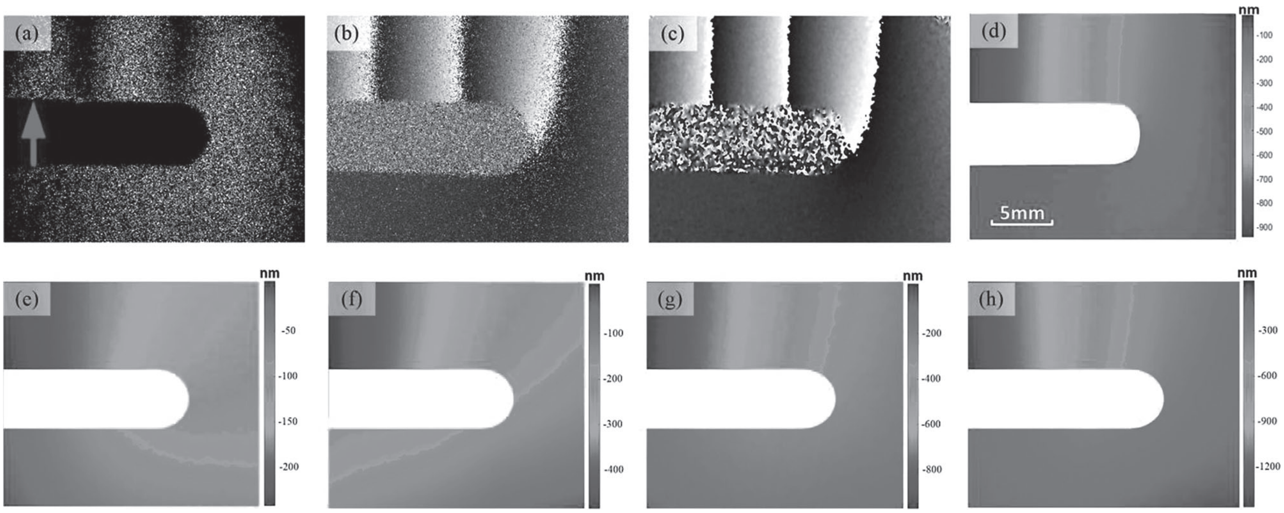
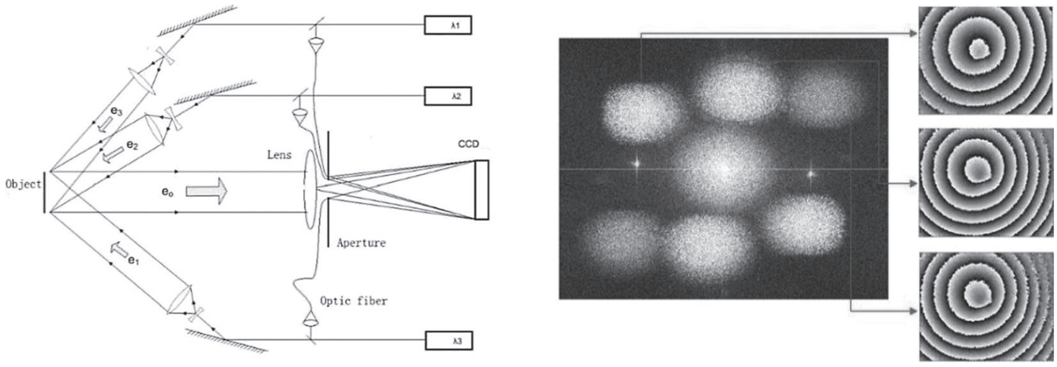


FIGURE 5.18 In-plane deformation measurement under a continuous loading by spatial phase dual beam illumination, (a) to (d) real-time fringe, original phase map, the filtered phase map and in-plane deformation, respectively, (c) to (h) selected in-plane deformations at different loading levels (from small to big).



(a) (b)

FIGURE 5.19 (a) Schematic diagram for simultaneous measurement of 3D deformations. (b) Spectrum of an image containing three speckle patterns with different spatial frequency shifts and the corresponding phase maps evaluated.

frequency shifts are built into each speckle pattern. The three speckle patterns from the three ESPI systems are simultaneously and synchronously recorded by the CCD camera into a single image. After applying an FT on the image with the three speckle patterns overlapped, a spectrum of the image with three spatial frequency shifts is obtained. A careful adjustment of reference beams is necessary so that the spectrum of each speckle pattern as described in Figure 5.9b will be separated in the total spectrum image. Figure 5.19b shows the total spectrum image which consists of three single spectrums corresponding to each spatial phase-shift ESPI set-up. In the figure, three single spectrums are oriented 45°, 90° and 135°, respectively. After the Windowed Inverse Fourier Transform, three phase maps corresponding to three directional measurements are evaluated (see the right side of Figure 5.19b).

As described above as well as in Equation (5.4), three ESPI set-ups generate three equations, and the deformation components *u, v* and *w* can be determined, thus enabling the simultaneous measurement of 3D deformations by recording only a single image. In combination with a high-speed camera and high-powered lasers, this function enables recording 3D deformation history/procedure under a dynamic loading, which has been demonstrated in Figure 5.20.

Other applications of spatial phase ESPI systems for measuring 3D deformations can be found in references [57–59].

5.4.3 Applications for Dynamic Measurements

Dynamic measurements can be classified into two categories: harmonic vibration measurement and transient measurement. The time-Averaged and stroboscopic-illumination methods are usually utilized for harmonic vibration measurement, while double-pulse spatial phase-shift ESPI is used for deformation measurement under a transient loading or an impact loading.

5.4.3.1 Time-Averaged Method with a Refreshed Reference Frame

When an object is excited by a harmonic vibration, the surface of the sample exhibits a characteristic amplitude response to the excitation frequency. If the vibration frequency is significantly higher than the acquisition rate of the CCD camera, the

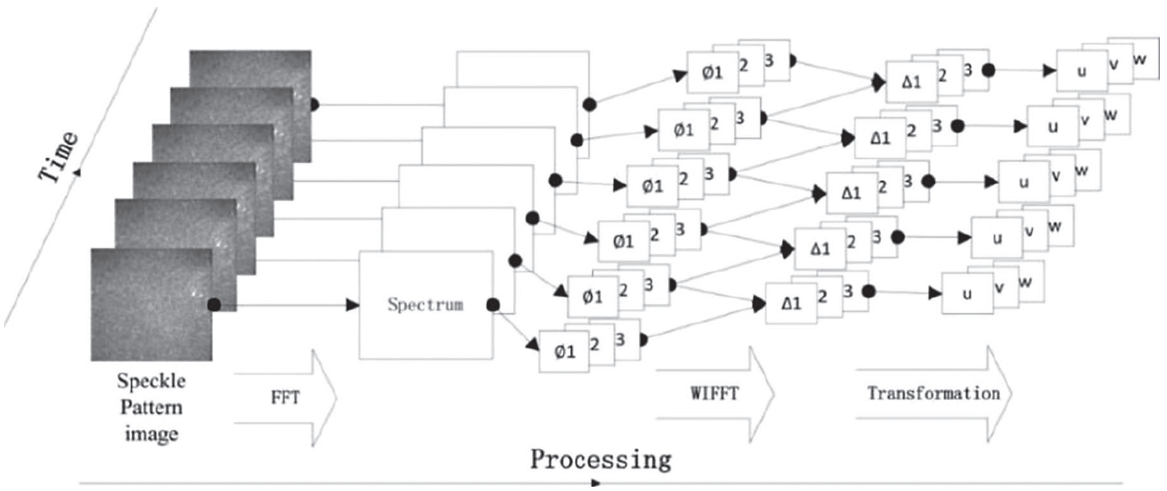


FIGURE 5.20 Principle and the image processing for recording and evaluating 3D deformation history under a dynamic loading.

resulting intensity in the camera plane is given by an averaged intensity value:

$$I_{\text{avg}} = a + b\cos\varphi J_0(\Omega) \tag{5.30}$$

where $J_0(\Omega)$ is the zero-order Bessel function of the first kind and $W = 4\pi w/\lambda$, where w is the out-of-plane deformation obtained by using an out-of-plane ESPI system. If a dual-beam in-plane ESPI system is used, in-plane deformation u or v will be measured and w in the equation above will be replaced by u or v. Equation (5.30) shows that the time-averaged intensity I_{avg} is modulated by the zero-order Bessel function. However, the fringes are not visible due to the strength of the self-interference term (bias term) "a". A typical procedure to remove the bias term is to subtract the intensity, shown in Equation 5.1, while the object is not vibrating from the intensity while the object is vibrating as shown in Equation (5.30). Thus, the subtractive results show a visible fringe pattern, which can be displayed on the monitor at video rate, and the fringes are modulated by $1 - J_0(\Omega)$ [13,60].

In the section, the time-averaged method with a refreshed reference frame whose fringes are modulated by J_0 will be introduced. While measuring, the current frame whose intensity equation is as shown in Equation (5.30) is always subtracted from the one that immediately precedes it rather than from the reference frame recorded at the stationary state. In addition, a phase shift of 180° performed by PZT mirror set in the path of the reference beam is introduced by every other frame during the video sequence. Now the subtraction of the intensities between these two frames generates:

$$\begin{aligned} I_{\text{avg-sub}} &= [a + b\cos\varphi J_0(\Omega)] - [a + b\cos\left(\varphi - 180^\circ\right) \\ J_0(\Omega)] &= 2b\cos\varphi J_0(\Omega) \end{aligned} \tag{5.31}$$

This time-averaged method with a refreshed reference frame has two significant advantages compared with the conventional time-averaged method: (i) The fringe pattern is modulated by $J_0(\Omega)$ rather than $[1 - J_0(\Omega)]$, thus the contrast of the fringes becomes much better than conventional method with a fixed reference frame, and (ii) the time between the acquisition of current and reference frame is greatly reduced, so the low-frequency noise caused by thermal air current and by low-frequency interruptions are greatly suppressed. This enables measurements to be performed without setting object on a vibration isolation table. Figure 5.21 shows the first six resonance frequencies and the mode shapes of an aluminium plate with the dimensions of $177 \times 127 \times 1.1\ \text{mm}^3$ clamped at one side (at the bottom). The measurements were performed on a table without vibration isolation facilities. The white lines on the images are the positions where the displacement is equal to zero, i.e. the node lines. The technique is usually used for qualitative analysis such as qualitative modal analysis and NDT by vibration excitation.

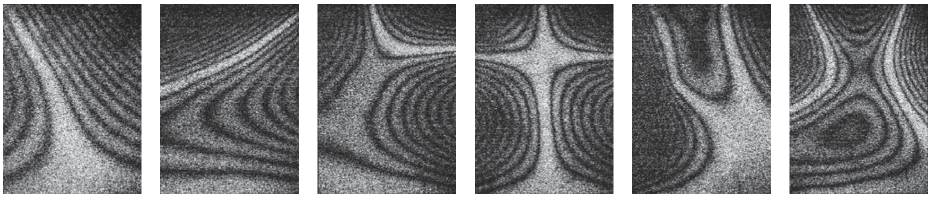
5.4.3.2 Stroboscopic Method

The time-averaged technique can be used to monitor the spatial distribution of the vibration mode and can also be used to detect subsurface defects and structural properties of the object [37]. Being mainly a qualitative technique, the extended procedure using stroboscopic illumination was developed for the quantitative measurement of phase distribution. When the object is harmonically excited, a stroboscopic ESPI technique could be applied to reveal the dynamic vibration behaviour and determine the mode shape. To generate stroboscopic illumination, a continuous wave laser is connected with an acoustic optical modulator which delivers short light pulses. A synchronizer and a control device are required for synchronizing the pulse with the vibration frequency to illuminate the object at each cycle of the vibration at a fixed position and control the pulse position and width.

Figure 5.22 shows the principle of repetitive stroboscopic acquisition for two trigger positions. The fundamental of the stroboscopic method is to freeze the deformation at the same state within multiple vibration periods by synchronized stroboscopic illumination while the camera takes an image (speckle pattern). Once the first image has been taken at the first trigger position, the light pulses are shifted to another position relative to the excitation, and another speckle pattern is captured. In fact, one or more images can be recorded at each trigger position, and both temporal and spatial phase-shift methods can be applied in this technique.

A subtraction between two recorded intensity images or two calculated phase distributions by the phase-shift technique creates a fringe pattern or a phase map depicting a deformation between the two states. The result could be an out-of-plane or in-plane deformation of a harmonically excited object, depending on which set-up is used. Figure 5.23 shows a vibration mode of an impeller at a frequency of 1708 Hz obtained by the stroboscopic illumination method in combination with the temporal phase-shift technique.

Other applications of the stroboscopic method for vibration measurement can be found in references [13.17,61].



126Hz 266Hz 517Hz 607Hz 694Hz 836Hz

FIGURE 5.21 The first six resonance frequencies and mode shapes for a plate clamped at one side (at the bottom) with dimensions of $127 \times 177 \times 1.1\ \text{mm}^3$.

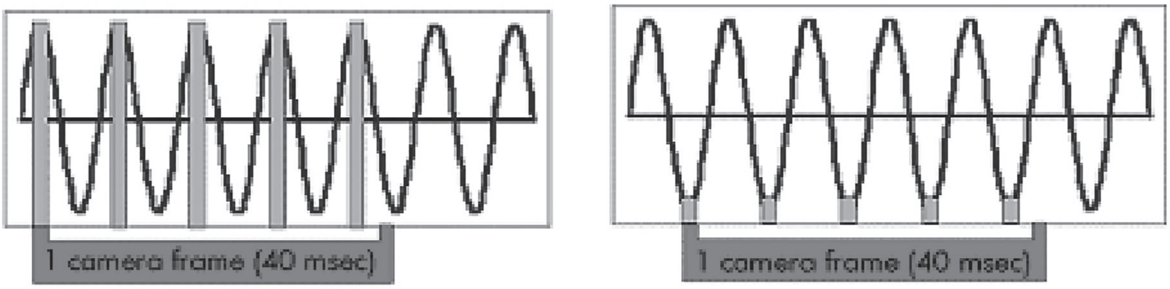


FIGURE 5.22 Principle of the stroboscopic method with an example of two trigger positions.

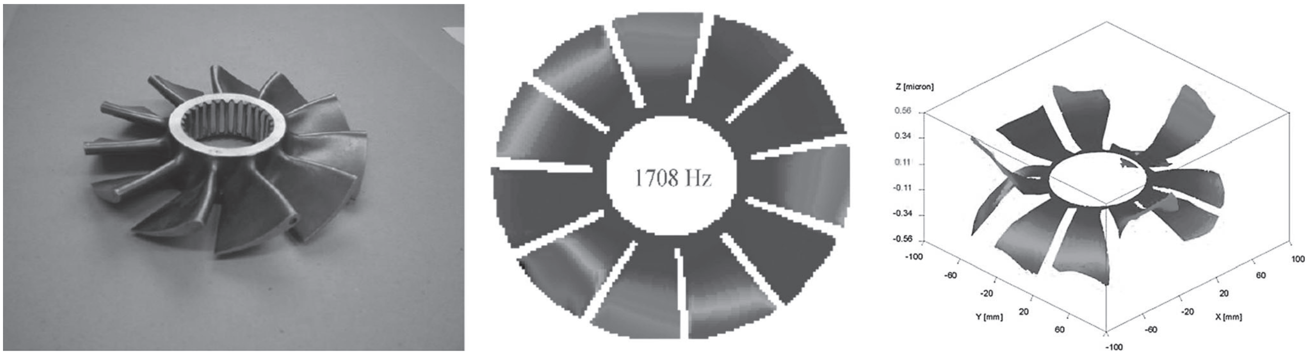


FIGURE 5.23 Vibration measurement of an impeller by the stroboscopic illumination method in combination with temporal phase-shift technique; left to right: image of the impeller, vibration mode at 1708 Hz and the 3D-display of the vibration mode.

5.4.3.3 Double-Pulse Method

The time-averaged method is suited well for qualitative vibration analysis, while the stroboscopic illumination method is well suited for quantitative vibration measurement if the object is under a harmonic excitation. In the case that the object is not excited harmonically or, more generally, in a transient way, both the time-averaged and stroboscopic illumination techniques no longer apply. A solution is given by the so-called pulsed ESPI technique. The object movement is literally frozen due to the short exposure time of a pulsed laser to illuminate the object, and the measurement resolves the deformation during the highly dynamic process. A pulsed Ruby laser illuminates the sample being tested. For reference purposes, a small part of the laser light is coupled to the image acquisition sensor via a glass fibre. Inside the sensor, the reference light and the light reflected from the object are superimposed (interfere) on a CCD camera. The interference patterns of two different object-loading situations are compared and reveal information about the deformation change of the object during loading. The delay of the two lasers typically can be varied between 2 and 800 micro-seconds (μs). This allows measurements of the change in object deformation (for example, due to an impact or a transient loading) for even highly dynamic processes. Different applications of double-pulse ESPI have been reported [62–64]. With the fast development of computer and camera techniques, using an extremely high-speed camera and high-power laser, dynamic measurements of deformation history, which is unable to be measured by the double-pulse laser technique, under a transient or impact loading using the spatial phase-shift ESPI should be reported in the near future.

5.4.3.4 Further applications of ESPI

Besides the applications in deformation measurement, NDT and vibration analysis described above, there are many other applications of ESPI both in engineering and across most scientific disciplines. In combination with the endoscopic, fibre optical and MEMS technologies [65], ESPI will enable the measurement of internal mechanical properties of an organ, which will open vast and immeasurable vistas for medical and biotic industries [66,67]. ESPI has also expanded its application into measuring material properties and many branches of experimental mechanics [48]. ESPI looks likely to become a standard tool for structural and material property identification, modelling assessment, residual stress, rheology, damage, and ageing in particular fields. It still has the potential for further exploration and applications [68].

5.5 Conclusions

This chapter has presented the state-of-the art techniques of ESPI. Particular emphasis has been given to three areas: principle of different ESPI systems; recent developments in temporal and spatial phase-shift techniques to extract the phase information and different applications of temporal and spatial phase-shift ESPI systems for deformation measurement under either static and dynamics loading and for NDT and vibration measurements. Continuous developments and improvements in software programs, computer technology and optical measuring techniques will enable additional enhancements of the basic ESPI instrument to be generated in the future.

REFERENCES

1. Rigden, J. D., & Gordon, E. I. 1962. The granularity of scattered optical maser light, *Proc. IRE*., 50: 2367–2368.
2. Butters, J. N., & Leendertz, J. A. 1971. A double exposure technique for speckle pattern interferometry, *J. Phys. E: Sci. Inst*., 4(4): 277.
3. Løkberg, O. J., & Høgmoen, K. 1976. Vibration phase mapping using electronic speckle pattern interferometry, *Appl. Opt*., 15(11): 2701–2704.
4. Butters, J. N., Jones, R., & Wykes, C. 1978. Electronic speckle pattern interferometry, In: *Speckle Metrology*. (A80-26519 09-35) New York, Academic Press, Inc., p. 111–158.
5. Creath, K. 1988. V phase-measurement interferometry techniques, *Prog. Opt*., 26: 349–393.
6. Rastogi, P. K. 2001. *Digital Speckle Pattern Interferometry and Related Techniques*, by PK Rastogi (Ed.), John Wiley & Sons Inc, United States, ISBN: 978-0-471-49052-4.
7. Yang, L., Zhang, P., Liu, S., Samala, P. R., Su, M., & Yokota, H. 2007. Measurement of strain distributions in mouse femora with 3D-digital speckle pattern interferometry, *Opt. Lasers Eng*., 45(8): 843–851.
8. Starukhin, P. Y., Kharish, N. A., Ulyanov, S. S., Lepilin, A. V., & Tuchin, V. V. 1997. Speckle techniques for blood microcirculation monitoring in periodontal treatment, *Opt. Diagnost. Biol. Fluids Adv. Techniq. Anal. Cytol*., 982: 299–308.
9. Yang, L., Steinchen, W., Schuth, M., & Kupfer, G. 1995. Precision measurement and nondestructive testing by means of digital phase shifting speckle pattern and speckle pattern shearing interferometry, *Measurement*, 16: 149–160.
10. Yang, L., & Xie, X. 2016. Digital Shearography: New developments and applications. *Measure of the Mechanics of Components*: 22–28, SPIE Press, Washington, USA.
11. Leendertz, J. A. 1970. Interferometric displacement measurement on scattering surface utilizing speckle effect, *J. Phys. E: Sci. Inst*., 3(3): 214–218.
12. Zhao, Q., Dan, X., Sun, F., Wang, T., Wu, S., Yang, L. 2018, Digital shearography for NDT: phase measurement technique and recent development, *Applied Sciences*, 8(12): 2662.
13. Steinchen, W., & Yang, L. 2003. *Digital Shearography: Theory and Application of Digital Speckle Pattern Shearing Interferometry*, SPIE Press, Bellingham, WA.
14. Schmit, J., Creath, K., & Kujawinska, M. 1993. Spatial and temporal phase-measurement techniques: a comparison of major error sources in one dimension, *Interferometry: Tech. Anal*., 1755: 202–212.
15. Yamaguchi, I. 2006. Chapter 5, Phase-shifting digitalholography, 145–171, In: *Digital Holography and Three-Dimensional Display-Principles and Applications*, by Ting-Chung Poon (Ed.) Springer, New York, USA, ISBN 978-0-387-31397-9.
16. Servin, M., & Cuevas, F. J. 1995. A novel technique for spatial phase-shifting interferometry, *J. Mod. Opt*., 42(9): 1853–1862.
17. Creath, K. 1985. Phase-shifting speckle interferometry, *Appl. Opt*., 24(18): 3053–3058.
18. Yang, L. X., Schuth, M., Thomas, D., Wang, Y. H., & Voesing, F. 2009. Stroboscopic digital speckle pattern interferometry for vibration analysis of microsystem, *Opt. Lasers Eng*., 47(2): 252–258.
19. Zhu, L. Q., Wang, Y. H., Xu, N., Wu, S. J., Dong, M. L., & Yang, L. X. 2013, Real-time monitoring of phase maps of digital shearography, *Opt. Eng*., 52(10): 101902.
20. Kao, C. C., Yeh, G. B., Lee, S. S., Lee, C. K., Yang, C. S., & Wu, K. C. 2002. Phase-shifting algorithms for electronic speckle pattern interferometry, *Appl. Opt*., 41(1): 46–54.
21. Creath, K. 1990. Phase-measurement techniques for nondestructive testing. In *Proceedings of SEM Conference on Hologram Interferometry and Speckle Metrology*, Baltimore, MD: 473–478.
22. Macy, W. W. 1983. Two-dimensional fringe-pattern analysis, *Appl. Opt*., 22(23): 3898–3901.
23. Wu, S., Zhu, L., Feng, Q., & Yang, L. 2012. Digital shearography with in situ phase shift calibration, *Opt. Lasers Eng*., 50(9): 1260–1266.
24. Rubin, L. F., & Wyant, J. C. 1979. Energy distribution in a scatterplate interferometer, *J. Opt. Soc. Am. A*, 69(9): 1305–1308.
25. Yang, L., Chen, F., Steinchen, W., & Hung, M. Y. 2004. Digital shearography for nondestructive testing: potentials, limitations, and applications, *J. Holography Speckle*, 1(2), 69–79.
26. Yang, L.X., & Ettemeyer, A. 2003. Strain measurement by 3D-electronic speckle pattern interferometry: potentials, limitation and applications, *Opt. Eng*., 42(5): 1257–1266.
27. Sirohi, R. S., Burke, J., Helmers, H., & Hinsch, K. D. 1997. Spatial phase shifting for pure in-plane displacement and displacement-derivative measurements in electronic speckle pattern interferometry (ESPI), *Appl. Opt*., 36(23): 5787–5791.
28. Baik, S.H., Park, S. K., Kim, C.J., & Kim S. Y. 2001. Two-channel spatial phase shifting electronic speckle pattern interferometer, *Opt. Commun*., 192(3–6): 205–211.
29. Burke, J., & Helmers, H. 1999. Performance of spatial vs. temporal phase shifting in ESPI, *Proc. SPIE*, 3744: 188–199.
30. Smythe, R., & Moore, R. 1984. Instantaneous phase measuring interferometry, *Opt. Eng*., 23(4): 234361.
31. Pedrini, G., Zou, Y. L., & Tiziani, H. J. 1996. Quantitative evaluation of digital shearing interferogram using the spatial carrier method, *Pure Appl. Opt.: J. Eur. Opt. Soc. Part A*, 5(3): 313.
32. Gao, X., Wu, S., & Yang, L. X. 2013. Dynamic measurement of deformation using Fourier transform digital holographic interferometry. In *Sixth International Symposium on Precision Mechanical Measurements*, 8916: 891617.
33. Gao, X., Yang, L. X., Wang, Y. H., Zhang, B. Y., Dan, X. Z., Li. J. R., & Wu, S. J. 2018. Spatial phase-shift dual-beam speckle interferometry, *Appl. Opt*., 57(3): 414–419.
34. Yang, L., Xie, X., Zhu, L., Wu, S., & Wang, Y. 2014. Review of electronic speckle pattern interferometry (ESPI) for

three-dimensional displacement measurement, *Chinese J. Mech. Eng*., 27(1): 1–13.

35. Ghiglia, D. C., & Pritt, M. D. 1998. *Two-Dimensional Phase Unwrapping: Theory, Algorithms, and Software*, Wiley, New York.
36. Kaufmann G. H. 1998. Unwrapping of electronic speckle pattern interferometry phase maps: evaluation of an iterative weighted algorithm *[J]. Opt. Eng*., 37: 622–628.
37. Yang, L. X., & Siebert, T. 2008. *Chapter 22, Digital Speckle Interferometry in Engineering from book: New Directions in Holography and Speckle* by Caulfield, H. J., & Vikram, C. S., (Eds), American Scientific Publishers. Stevenson Ranch, California, USA.
38. Paoletti, D., Spagnolo, G. S., Zanetta, P., Facchini, M., & Albrecht, D. 1994. Manipulation of speckle fringes for nondestructive testing of defects in composites, *Opt. Laser Technol*., 26(2): 99–104.
39. Jones, R., & Wykes, C. 1989. *Holographic and Speckle Interferometry*, 6. Cambridge University Press, Cambridge, England.
40. Stetson, K. A. 2015. Strain field measurement by transverse digital holography, *Appl. Opt*., 54(19): 6065–6070.
41. Fung, Y. C. 1994. *A First Course in Continuum Mechanics: for Physical and Biological Engineers and Scientists*, Prentice Hall, Englewood Cliffs, NJ.
42. Joenathan, C., Pfister, B., & Tiziani, H. J. 1990. Contouring by electronic speckle pattern interferometry employing dual beam illumination, *Appl. Opt*., 29(13): 1905–1911.
43. Peng, X., Diao, H., Zou, Y., & Tiziani, H. J. 1992. Contouring by modified dual-beam ESPI based on tilting illumination beams, *Optik*, 90(2): 61–64.
44. Ganesan, A. R., & Sirohi, R. S. 1989, January. New method of contouring using digital speckle pattern interferometry (DSPI). In *1988 Dearborn Symposium*: 327–332.
45. Takeda, M., & Yamamoto, H. 1994. Fourier-transform speckle profilometry: three-dimensional shape measurements of diffuse objects with large height steps and/or spatially isolated surfaces, *Appl. Opt*., 33(34): 7829–7837.
46. Stetson, K. A. 1990. Use of sensitivity vector variations to determine absolute displacements in double exposure hologram interferometry, *Appl. Opt*., 29(4): 502–504.
47. Siebert, T., Splitthof, K., & Ettemeyer, A. 2004. A practical approach to the problem of the absolute phase in speckle interferometry, *J. Holography Speckle*, 1(1): 32–38.
48. Jacquot, P. 2008. Speckle interferometry: a review of the principal methods in use for experimental mechanics applications, *Strain*, 44(1): 57–69.
49. Sun, L., Yu, Y., & Zhou, W. 2015. 3D deformation measurement based on colorful electronic speckle pattern interferometry, *Optik-Int. J Light Elect. Opt*., 126(23): 3998–4003.
50. Pouet, B. F., & Krishnaswamy, S. 1993. Additive/subtractive decorrelated electronic speckle pattern interferometry, *Opt. Eng*., 32(6): 1360–1370.
51. Wykes, C. 1982. Use of electronic speckle pattern interferometry (ESPI) in the measurement of static and dynamic surface displacements, *Opt. Eng*., 21(3): 213400–213400.
52. Moore, A. J., Hand, D. P., Barton, J. S., & Jones, J. D. 1999. Transient deformation measurement with electronic speckle pattern interferometry and a high-speed camera, *Appl. Opt*., 38(7): 1159–1162.
53. Løkberg, O. J. 1984. ESPI—the ultimate holographic tool for vibration analysis, *J. Acoustical Soc. Am*., 75(6): 1783–1791.
54. Fricke-Begemann, T., & Burke, J. 2001. Speckle interferometry: three-dimensional deformation field measurement with a single interferogram, *Appl. Opt*., 40(28): 5011–5022.
55. Wang, Y., Sun, J., Li, J., Gao, X., Wu, S., & Yang, L. 2016. Synchronous measurement of three-dimensional deformations by multi camera digital speckle patterns interferometry, *Opt. Eng*., 55(9): 091408–091408.
56. Gao, X., Wang, Y., Li, J., Dan, X., Wu, S., & Yang, L. 2017. Spatial carrier color digital speckle pattern interferometry for absolute three-dimensional deformation measurement, *Opt. Eng*., 56(6): 066107–066107.
57. Sjödahl, M., & Saldner, H. O. 1997. Three-dimensional deformation field measurements with simultaneous TV holography and electronic speckle photography, *Appl. Opt*., 36(16): 3645–3648.
58. Schedin, S., Pedrini, G., Tiziani, H. J., & Santoyo, F. M. 1999. Simultaneous three-dimensional dynamic deformation measurements with pulsed digital holography, *Appl. Opt*., 38(34): 7056–7062.
59. Gu, G., Wang, K., Wang, Y., & She, B. 2016. Synchronous triple-optical-path digital speckle pattern interferometry with fast discrete curvelet transform for measuring three-dimensional displacements, *Opt. Laser Technol*., 80: 104–111.
60. Lu, B., Yang, X., Abendroth, H., & Eggers, H. 1989. Time-average subraction method in electronic speckle pattern interferometry, *Opt. Commun*., 70(3): 177–180.
61. Zhu, L. Q., Wu, S. J., & Yang, L. X., 2014. Stroboscopic digital shearographic system for vibration analysis of large area object, *Inst Experiment. Techn*., 57(4): 493–498.
62. Eme, O., Guist, C., Bode, H., & Ettemeyer, A. 1998. 3D-ESPI for vibration analysis on catalytic converters. In *Proceedings of the SEM Spring Conference on Experimental Mechanics*: 345–347.
63. Van der Auweraer, H., Steinbichler, H., Vanlanduit, S., Haberstok, C., Freymann, R., Storer, D., & Linet, V. 2002. Application of stroboscopic and pulsed-laser electronic speckle pattern interferometry (ESPI) to modal analysis problems, *Measure. Sci. Technol*., 13(4): 451.
64. Pedrini, G., Pfister, B., & Tiziani, H. 1993. Double pulse-electronic speckle interferometry, *J. Modern Opt*., 40(1): 89–96.
65. Murukeshan, V. M., & Sujatha, N. 2007. All fiber based multispeckle modality endoscopic system for imaging medical cavities, *Rev. Sci. Inst*., 78(5): 053106.
66. Kemper, B., Dirksen, D., Avenhaus, W., Merker, A., & von Bally, G. 2000. Endoscopic double-pulse electronic-speckle-pattern interferometer for technical and medical intracavity inspection, *Appl. Opt*., 39(22): 3899–3905.
67. Zaslansky, P., Currey, J. D., Friesem, A. A., & Weiner, S. 2005. Phase shifting speckle interferometry for determination of strain and Young's modulus of mineralized biological materials: a study of tooth dentin compression in water, *J. Biomed. Opt*., 10(2): 024020–02402013.
68. Chen, F., Luo, W. D., Dale, M., Petniunas, A., Harwood, P., & Brown, G. M. 2003. High-speed ESPI and related techniques: overview and its application in the automotive industry, *Opt. Lasers Eng*., 40(5): 459–485.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

6

Optical Fibre Hydrophones

Geoffrey A. Cranch and Philip J. Nash

CONTENTS

6.1 Introduction 82
6.2 Basic Principles 82
    6.2.1 Sonar System Requirements 82
    6.2.2 Brief History 83
    6.2.3 Interferometric Hydrophone Basic Principles 83
6.3 Optical Fibre Interferometry 84
    6.3.1 Interferometer Configurations 84
    6.3.2 Lasers 85
    6.3.3 Modulation Properties 86
    6.3.4 Noise Properties 86
    6.3.5 Summary of Characteristics of Lasers Used in Interferometric Fibre Optic Sensors 87
    6.3.6 Components 87
6.4 Acoustic Interactions 88
    6.4.1 The Basic Transduction Mechanism 88
    6.4.2 Coated Fibres 89
    6.4.3 Mandrel Hydrophones 90
    6.4.4 Hydrophone Responsivity with Depth 92
    6.4.5 Hydrophone Response at Higher Frequencies and Directionality 93
    6.4.6 Sensitivity to Other Effects 93
    6.4.7 Practical Hydrophone Designs 93
6.5 Signal Processing 94
    6.5.1 Passive Interrogation Schemes 94
    6.5.2 Noise Sources and Phase Resolution 96
    6.5.3 Dynamic Range 97
    6.5.4 Digital Techniques 98
6.6 Optical Systems and Multiplexing 98
    6.6.1 The FDM Technique 98
    6.6.2 TDM Architectures 99
    6.6.3 TDM Architectures Analysis 100
    6.6.4 Large-Scale Array Architectures 101
    6.6.5 Overcoming Polarization-Induced Signal Fading 101
    6.6.6 Future Trends in Optical Hydrophone Technology—Fibre Laser Sensors 102
6.7 The Optical Geophone 102
    6.7.1 The Basic Transduction Mechanism 103
    6.7.2 Alternative Geophone Designs 103
    6.7.3 System Configurations for Geophone Use 104
6.8 Application Studies 104
References 106
Further Reading 109

6.1 Introduction

This chapter provides a description of two sensors used for underwater acoustic and seismic measurements: the optical fibre hydrophone, which is a device for the measurement of underwater acoustic pressure signals, and the optical fibre geophone, which measures seismic vibration. Optical fibre hydrophones form an interesting and important case study in the field of optical fibre sensors. There are several reasons for this. First, optical hydrophones were one of the very first optical fibre sensor technologies to be developed, in the late 1970s, and have a history of continuous development extending over more than 20 years. Second, because of this long history and significant level of investment, they are one of the most well-developed of all types of optical fibre sensor in terms of the performance achieved, the levels of multiplexing demonstrated and the harshness of the environments in which they have successfully operated. The hydrophone application is an area in which optical fibre sensors have demonstrated clear advantages over any other sensing approach. The development of the optical hydrophone has, therefore, had a significant impact on the development of optical fibre sensors, in general, and is a classic case study of a laser-based optical sensing system. More recently, the optical geophone, which relies on similar design principles to the hydrophone but utilizes a different mechanical transduction technique, has found application in the field of seismic surveying, principally in the oil and gas industry. The requirements here are for very large scale, highly multiplexed systems, often combining hydrophones and geophones, and since around the year 2000 some of the most significant developments in large-scale sensing systems have been used for this application.

This chapter is divided into seven sections. Section 6.2 gives a brief history of hydrophone development, describes hydrophone requirements and summarizes the various approaches to optical fibre acoustic measurement. Hydrophones based on interferometric techniques are shown to achieve the best performance and this technology forms the subject of the remaining sections. Section 6.3 outlines the basic principles of optical fibre interferometry, discussing in particular the central role of the laser. Section 6.4 describes the basic transduction mechanism and outlines the various approaches to hydrophone design. Section 6.5 discusses the various interrogation techniques used to extract the acoustic information from the optical signal and Section 6.6 extends this to multiplexed systems and discusses new optical fibre sensor-based hydrophone technologies. Section 6.7 describes the optical geophone and associated technologies required for the seismic industry. Finally, Section 6.8 presents several application studies to show how the technology has been applied to real systems that have been tested in the field, in both the defence and oil and gas areas.

6.2 Basic Principles

6.2.1 Sonar System Requirements

Hydrophones have a wide range of applications in underwater metrology. For instance, they are used for the tracking of marine wildlife, as part of bathymetric survey systems, and in seismic surveying for new oil fields. However, an important application for hydrophones remains military sonar systems and it is this that has provided the driving force for the development of optical fibre hydrophones.

The primary purpose of military sonar is the detection of submarines. These sonar systems are typically attached to submarines or surface vessels, although seabed arrays are also used. Platform-mounted sonar includes hull-mounted systems directly attached to the vessel or towed arrays that comprise long flexible arrays pulled behind the vessel. In both cases, the size and complexity of the arrays have been steadily increasing, in an attempt to detect very quiet submarine targets. Modern systems can comprise up to several thousand individual hydrophone channels that are processed to discriminate against background noise. Existing systems typically use electroceramic materials such as piezoelectric or electrostrictive materials [1]. These sensors, which have been very well developed, achieve adequate performance but, as system size increases, their use presents a number of problems. In order to minimize on numbers of individual cables, such arrays require a high level of multiplexing, and for electroceramic materials, this requires a considerable amount of electronics within the arrays. This increases the weight, complexity and cost of the arrays and renders them prone to failure. Such arrays also require considerable amounts of electrical power to be supplied to the array.

The optical fibre hydrophone attracted interest within the military community because it offered the potential of a simple, lightweight, robust sensor that could be multiplexed and remotely interrogated without any need for electronics or electrical power within the array. The requirement is, therefore, to develop a hydrophone that achieves performance at least as good as existing technologies but with the previously outlined advantages. Table 6.1 sets out the range of likely performance requirements for most practical systems.

Three performance requirements are particularly worthy of note. First, since the requirement is for the measurement of a dynamic quantity (i.e. acoustic), the system must have a broadband capability and low-frequency drift (less than ~1 Hz) due to changes in ambient pressure and temperature that can be ignored. However, the optical hydrophone can, in

TABLE 6.1

Hydrophone Performance Requirements

Parameter	Typical Requirement
Frequency range	5–10 000 Hz
System noise floor	Equivalent to deep sea state zero after [2] (40 dB re μPa Hz$^{-1/2}$ at 500 Hz)[a]
Dynamic range	120 dB
Number of hydrophones	100–10 000
Sensor–sensor cross-talk	less than –40 dB
Operating depth	500 m
Survival depth	1000 m

[a] Ocean ambient acoustic noise is frequency-dependent. Deep sea state zero can be approximated by the formula 90–16.65 log f dB re 1 μPa Hz$^{-1/2}$, where f is the acoustic frequency in Hz. This is equivalent to an $f^{-0.83}$ dependence on frequency.

principle, reliably operate down to near dc. In practice, the low-frequency cut-off is set in the sensor demodulation electronics. Second, since the sensitivity of bare fibre to pressure is low, considerable amplification of the effect of pressure is required along with high-resolution interrogation in order to achieve the acoustic resolution stated in Table 6.1. Indeed, it will be shown that a strain resolution in 100 m of fibre of 10^{-14} is achievable! Finally, the number of hydrophones required implies a high level of multiplexing, perhaps in excess of 100 hydrophones per fibre. These demanding requirements will dictate the choice of sensing approach.

Other important requirements relate to the practical design of the sensor. For example, the hydrophone should exhibit an omni-directional response over the acoustic frequency range of interest, which implies dimensions much smaller than an acoustic wavelength. It should also have a flat frequency response over the operating frequency range of interest, should tolerate high static pressures, exhibit excellent signal selectivity and be robust enough to survive in the ocean environment.

6.2.2 Brief History

The first reported studies into optical fibre hydrophones date from 1977 when Bucaro et al. [3,4] reported that the pressure changes associated with an acoustic field produced measurable changes within a length of optical fibre. A wide range of sensing approaches based on both intrinsic and extrinsic sensors were investigated [5]. In an intrinsic sensor, the acoustic signal directly affects some property of the fibre, while in an extrinsic sensor, the fibre acts as a means for supplying light to a separate sensing element. All the sensors investigated can be classed as one of three types depending on which property of light they affect: amplitude [6], polarization [7] or phase. However, it became apparent that interferometric sensors based on phase detection offered both the highest sensitivity and the greatest potential for multiplexing, so with few exceptions, the interferometric approach became the preferred choice. This approach was initially developed in parallel at a number of locations, including the Naval Research Laboratory (NRL) [8] and Stanford University in the US [9] and by Plessey Naval Systems in the UK [10]. Interest in the technology continued to develop, and by 1998, there were at least ten countries worldwide working on interferometric hydrophone systems. Important milestones in the development of the technology include the first at-sea demonstration of an array in 1986 [11], the first at-sea deployment of a towed array in 1990 [12] and the first deployment of a large (64 hydrophone) seabed array in 1996 [13].

6.2.3 Interferometric Hydrophone Basic Principles

When a single-mode optical fibre is located within an acoustic field, the pressure and temperature fluctuations associated with the wave propagation generate strain within the optical fibre. The total phase of an optical beam propagating through the fibre will, thus, be modulated in proportion to the magnitude of the pressure and temperature fluctuations. This forms the basic principle of the interferometric hydrophone. By incorporating the fibre into one arm of an interferometer, this phase shift can be converted into an intensity modulation at an optical receiver.

By considering a length of fibre, L, the total phase of the optical signal propagating through the fibre, ϕ, is given by

$$\phi = nkL \tag{6.1}$$

where n is the effective refractive index of the fibre, which can be roughly approximated to the refractive index of the fibre core, k is the free space propagation constant equal to $2\pi/\lambda_0$ and λ_0 is the optical wavelength in a vacuum. Differentiation of (6.1) and rearranging yield

$$\frac{\Delta\phi}{\phi} = \frac{\Delta L}{L} + \frac{\Delta n}{n} + \frac{\Delta k}{k}. \tag{6.2}$$

Thus, the phase of the optical signal is affected by three factors: (i) a physical length change of the fibre, ΔL; (ii) a change in refractive index of the fibre Δn, due to the stress-optic effect and (iii) a change in the wavelength of the input signal, Δk. The first two effects provide a means of transduction of an acoustic signal into optical phase, and the magnitude of these effects is described in Section 6.4. The third term provides a means of passive interrogation of a fibre interferometer by modulation of the input wavelength and is discussed in Section 6.5.

To illustrate how a phase change can be measured interferometrically, we now analyze the fibre Mach–Zehnder interferometer, shown in Figure 6.1, which forms the simplest interferometric sensor arrangement.

The sensor arrangement comprises a laser, a twin-arm interferometer and a photodetector (usually a photodiode). The optical circuit is assembled using a single-mode fibre to prevent modal interference. The light from the source is split into the two arms of the interferometer using a fibre-directional coupler (a fibre analogue of the free-space beamsplitter), such that equal power is launched into each arm and recombined using a second coupler. The hydrophone consists of a length of optical fibre forming one arm of the interferometer that is exposed to the acoustic field to be measured, while the delay coil in the other arm is shielded from the acoustic field. For the hydrophone to exhibit an omni-directional response (i.e. the hydrophone response has no dependence on direction of signal field), the hydrophone dimensions must be much smaller than the acoustic wavelength. If the launched power is $2P$, then the optical fields incident on the photodiode can be expressed as

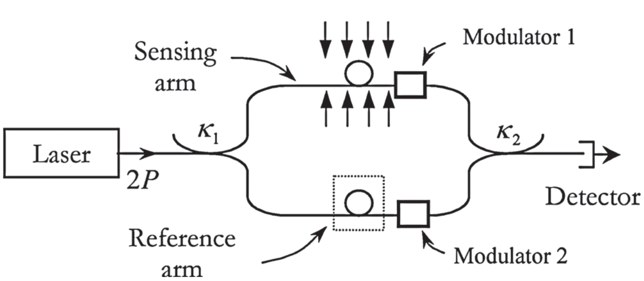


FIGURE 6.1 Basic Mach–Zehnder fibre-optic interferometric sensor arrangement.

$$E_s = \sqrt{2P\alpha_s k_1 k_2} \cdot e^{i(\omega_s t+\phi_s)} \tag{6.3}$$

$$E_r = \sqrt{2P\alpha_r(1-k_1)(1-k_2)} \cdot e^{i(\omega_r t+\phi_r)} \tag{6.4}$$

where α_s and α_r represent optical loss associated with the signal and reference arms, respectively, κ_1 and κ_2 are the power coupling coefficients of the directional couplers, ϕ_s and ϕ_r are the total phase of the signals from the sensing and reference arms, respectively, and ω_s and ω_r are the optical frequencies. The total optical field is thus $E_{tot} = E_s + E_r$, and the photocurrent is given by

$$i_{ph} = r\langle E_{tot} cc E_{tot} \rangle \tag{6.5}$$

where *cc* denotes the complex conjugate and $\langle\ \rangle$ indicates a time average over the detector time constant. r is the photodiode responsivity in A W^{-1} and $r = e\eta/h\nu$ where e is the electron charge, η is the photodiode quantum efficiency (ratio of the number of electron–hole pairs generated for one photon absorption) and h is Planck's constant. Evaluating equation (6.5) with $\Delta\omega = \omega_s - \omega_r$ yields

$$\begin{aligned} i_{ph} = 2\Pr\big[&\alpha_s\kappa_1\kappa_2 + \alpha_r(1-\kappa_1)(1-\kappa_2) \\ &+ 2\sqrt{\alpha_s\alpha_r\kappa_1\kappa_2(1-\kappa_1)(1-\kappa_2)}\cos(\Delta\omega t + \phi_s - \phi_r)\big]. \end{aligned} \tag{6.6}$$

Thus, the amplitude of the interference term is dependent on the optical losses in the interferometer and the coupler split ratios. It will also be dependent on the coherence properties of the laser and the birefringence properties of the interferometer arms. For this analysis, we shall include these effects by introducing a fringe visibility factor, V, where $0 \le V \le 1$ and return to the cause of these effects later. In practice, it is possible to reduce the optical losses to near zero, hence $\alpha_s \approx \alpha_r \approx 1$, and the interference term is maximized by setting $\kappa_1 = \kappa_2 = 1/2$, thus equation (6.6) reduces to

$$i_{ph} = rP + rPV\cos(\Delta\omega t + \phi_s - \phi_r) \tag{6.7}$$

where $\Delta\omega = \omega_s - \omega_r$. In the arrangement shown in Figure 6.1, $\omega_s = \omega_r$, thus $\Delta\omega = 0$, and in this case, the detection regime is referred to as *homodyne*. However, if the optical frequencies were shifted relative to one another (using a modulator) such that $\Delta\omega > 0$, the detection regime is referred to as *heterodyne*. This can be used as a means of remotely interrogating multiplexed interferometers and the techniques available to achieve this frequency shift are discussed in Sections 6.3 and 6.5.

For the case when $\Delta\omega = 0$, the output of the fibre interferometer will be proportional to the cosine of the phase difference between the received signals. Environmental effects on the interferometer arms due to changes in ambient temperature and pressure will cause a slowly varying phase drift that will prevent a higher frequency signal phase term from being linearly recovered. This is known as the fading problem in fibre interferometers and a passive signal recovery technique is thus required to extract the phase signal of interest. It is also shown in equation (6.7) that an ambiguity in the photocurrent occurs when the phase shift corresponding to a signal is over 2π radians. Some form of fringe-counting technique is, therefore, required to measure large phase shifts. These will both be discussed in Section 6.5.

6.3 Optical Fibre Interferometry

6.3.1 Interferometer Configurations

The Mach–Zender interferometer described earlier was the basis for many of the early hydrophone systems and is still used in some present designs. However, a number of other interferometer configurations offer various advantages for practical systems and are often used in more recent systems. As well as the Mach–Zender interferometer, two other interferometers that have been used to implement the fibre hydrophone are shown in Figure 6.2. These are the Michelson and the Sagnac interferometers. It is instructive to compare interferometers by determining the frequency response of each. For this, we shall use the approach presented in [14] and [15].

If a harmonically varying acoustic signal with no spatial dependence is incident on an optical fibre of length, L_s, this will cause a perturbation of the propagation constant, β, of the optical signal travelling in the fibre. This is due to the effects illustrated in equation (6.2) (i.e. the length and refractive index change). The propagation constant can, thus, be expressed as

$$\beta(x,t) = \beta_0 + \delta\beta e^{i(\omega_a t)} \tag{6.8}$$

where $\beta_0(= nk)$ is the unperturbed propagation constant, $\delta\beta$ is the amplitude of the propagation constant modulation, ω_a is the angular acoustic frequency and x is the axial coordinate along the fibre. The total phase of the optical signal, $\phi_s(t)$, is then given by integration of the propagation constant over the sensor fibre length, L_s,

$$\phi_s(t) = \int_0^{L_s} \beta\left(x, t - \frac{xn}{c}\right) dx \tag{6.9}$$

where c is the light velocity in a vacuum. We shall assume in the following analysis that only the sensor fibre is exposed to the acoustic signal.

Mach–Zender. The frequency response of the Mach–Zender interferometer, shown in Figure 6.1, can be determined by substituting equation (6.8) into equation (6.9) to obtain the responsivity,

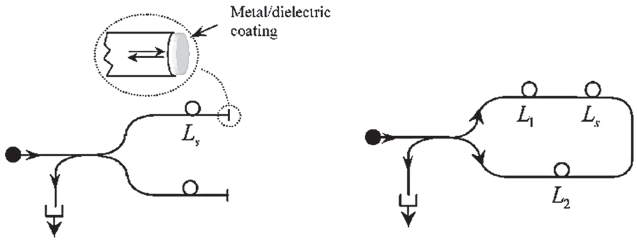


FIGURE 6.2 Interferometer configurations: (a) Michelson and (b) Sagnac.

$$|\Delta\phi_{\mathrm{MZ}}| = \delta\beta L_{\mathrm{s}}\left|\sin c\left(\frac{\omega_{\mathrm{a}}n}{2c}L_{\mathrm{s}}\right)\right| = \frac{2c\delta\beta}{\omega_{\mathrm{a}}n}\left|\sin\left(\frac{\omega_{\mathrm{a}}n}{2c}L_{\mathrm{s}}\right)\right|. \tag{6.10}$$

which at all practical acoustic frequencies can be approximated to,

$$|\Delta\phi_{\mathrm{MZ}}| \approx \delta\beta L_{\mathrm{s}} \tag{6.11}$$

Michelson. The Michelson interferometer is shown in Figure 6.2a. Light from a laser source is split into two fibres by a directional coupler. At the end of each fibre arm, a mirror coating is applied to the end of the cleaved fibre. To achieve high reflectivity, the fibres must be accurately cleaved before the mirror coating is applied by either chemical or vacuum deposition. The light is, thus, reflected back along the same path to exit through the remaining port of the directional coupler. Care must be taken to avoid optical feedback into the laser cavity by use of an optical isolator on the input port of the directional coupler.

The frequency response for the Michelson interferometer is determined in a similar way to that shown earlier and the phase responsivity is given by

$$\begin{aligned}|\Delta\phi_{\mathrm{M}}| &= 2\delta\beta L_{\mathrm{s}}\left|\sin c\left(\frac{\omega_{\mathrm{a}}n}{2c}L_{\mathrm{s}}\right)\cos\left(\frac{\omega_{\mathrm{a}}n}{2c}L_{\mathrm{s}}\right)\right| \\ &= \frac{2c\delta\beta}{\omega_{\mathrm{a}}n}\left|\sin\left(\frac{\omega_{\mathrm{a}}n}{c}L_{\mathrm{s}}\right)\right|\end{aligned} \tag{6.12}$$

which can be approximated to

$$|\Delta\phi_{\mathrm{M}}| \approx 2\delta\beta L_{\mathrm{s}}. \tag{6.13}$$

Sagnac. The use of the Sagnac interferometer as a hydrophone was first reported in [16], and the basic interferometer is shown in Figure 6.2b. Here, the light is split into two paths by a directional coupler that travels in opposite directions through the same closed loop. The loop contains two delay coils of length L_1 and L_2 and a sensor coil of length L_{s}. A static perturbation acting on the complete loop will have no net effect on the phase delay between the two beams, because both beams will be affected equally. However, a dynamic signal acting on the sensor but not on the delay coils (i.e. non-symmetrically about the loop) will introduce a phase difference between the two counter-propagating beams due to the fact that one beam is in a different part of the loop when the other beam is passing through the sensor. Analysis of the Sagnac interferometer requires separate integration of the counter-propagating beams. This has been carried out in [14] where it was shown that

$$|\Delta\phi_{\mathrm{s}}| = 2\delta\beta L_{\mathrm{s}}\left|\sin c\left(\frac{\omega_{\mathrm{a}}n}{2c}L_{\mathrm{s}}\right)\sin\left(\frac{\omega_{\mathrm{a}}n}{2c}L_{\mathrm{d}}\right)\right| \tag{6.14}$$

where $L_{\mathrm{d}} = L_2 - L_1$. For the case when $\omega_{\mathrm{a}} \ll \pi c/nL_{\mathrm{d}}$ (i.e. below the first null in the phase responsivity), the responsivity of the Sagnac interferometer is reduced to

$$|\Delta\phi_{\mathrm{s}}| \approx 2\delta\beta\frac{\omega_{\mathrm{a}}n}{2c}L_{\mathrm{s}}L_{\mathrm{d}}. \tag{6.15}$$

It is now possible to compare the frequency response of the three interferometers. The normalized responsivity, $|\Delta\phi| / (\delta\beta L_{\mathrm{s}})$, of each interferometer is plotted in Figure 6.3 as a function of normalized frequency, $\omega_{\mathrm{a}}nL_{\mathrm{s}} / 2c$.

The responsivity of the Michelson interferometer is thus twice (6 dB) the responsivity of the Mach–Zender interferometer due to the dual pass of the optical signal through the sensor coil. The responsivity of the Sagnac interferometer is proportional to frequency for $\omega_{\mathrm{a}} \ll \pi c/nL_{\mathrm{d}}$. This property of the interferometer has been used to increase the available dynamic range in an optical hydrophone since the frequency dependence of the responsivity compensates for the background acoustic noise spectrum (which exhibits a $f^{-0.83}$ dependence on frequency), such that the resulting sensor noise floor is roughly flat [17]. However, in general for the hydrophone application, the delay coil length required is very large. For example, for a practical hydrophone with L_{s} = 100 m, the interferometer responsivity must be approximately equal to its Michelson equivalent at $f_{\mathrm{a}} \sim 10$ kHz. This requires a delay coil length, $L_{\mathrm{d}} = 100L_{\mathrm{s}}$, as shown in Figure 6.3, or 10 km to give the required responsivity.

6.3.2 Lasers

The laser is an important component in the interferometric fibre optic hydrophone and may determine the overall performance of the sensor. Indeed, to a large extent, the development of optical hydrophone systems has been dependent on the availability of lasers with the required performance. In general, the laser must exhibit high stability in terms of output power, frequency, polarization and wavelength and be required to maintain this stability in the presence of high ambient acoustic noise and vibration. It is thus necessary to characterize the laser in terms of low-frequency intensity noise (up to 1 MHz) and frequency noise. For most recent hydrophone systems, lasers operating in the 1310 or 1550 nm range are used. This ensures compatibility with optical communication grade components and in the case of 1550 nm allows the use of erbium-doped fibre amplifiers (EDFA). To allow high-resolution demodulation of the sensor signal, the laser must also maintain a single longitudinal mode output (i.e.

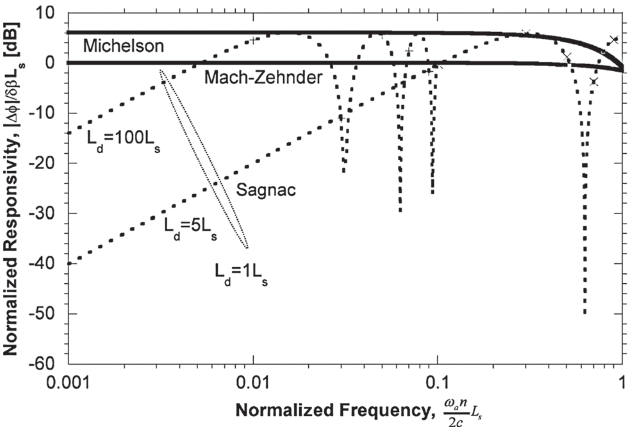


FIGURE 6.3 Normalized responsivity *versus* normalized frequency for the Mach–Zender, Michelson and Sagnac interferometers.

single frequency) and single polarization mode. Also, the laser output must be coupled into a single-mode fibre with relatively low loss (a point often taken for granted by end-users).

6.3.3 Modulation Properties

We first review the frequency modulation properties of lasers commonly used in optical hydrophone systems. It will be shown, in Section 6.4, how these modulation properties can be used to implement passive interrogation schemes for fibre interferometers to overcome the fading problem described earlier.

The semiconductor laser diode (LD) is a commonly used source in interferometric sensors. It can be designed to operate at 1310 or 1550 nm and produces in-fibre output powers typically up to 50 mW. Appropriate design of the laser ensures robust single longitudinal mode output and single polarization mode. The output of the laser is usually coupled into a single-mode polarization-maintaining fibre by butting a cleaved fibre to the laser junction. Lasing in the LD is achieved by injecting a current into the active region of the junction larger than the threshold current. The dependence of the laser emission frequency on the active region refractive index, and the Fabry–Pérot cavity length provides a means of frequency modulation. These two parameters are dependent on the drive current due to three effects: (i) the refractive index of the junction is a function of the charge-carrier density; (ii) the refractive index is a function of the junction temperature through the thermo-optic effect; and (iii) the Fabry-Perot cavity length is dependent on the junction temperature due to thermal expansion. The dependence of the optical emission frequency, v, on the drive current, i, can be characterized in terms of dv/di and the frequency dependence of this quantity has been investigated for several LDs [18]. It has been found that at low modulation frequencies, the magnitude of dv/di is dominated by the thermo-optic effect and at high frequencies by the charge-carrier density. Commercially available LDs based on GaAlAs were shown to exhibit near DC values of dv/di of between 3 and 11.5 GHz mA^{-1}.

The diode-pumped Nd:YAG laser is often used in high-performance hydrophone systems due to its excellent frequency stability. These lasers operate at 1310 nm and produce up to 200 mW of in-fibre power in a single longitudinal mode and single polarization mode. A piezoelectric element can be placed across the laser junction that allows a controlled acoustic wave to propagate across the junction. This provides a means of frequency modulation of the laser output through the acousto-optic effect. The magnitude of this effect is quantified in terms of dv/dV where V is the voltage across the piezoelectric element and has been measured to be about 2.8 MHz V^{-1} up to 100 kHz [19].

6.3.4 Noise Properties

Laser intensity noise. Laser intensity noise is caused by random fluctuations in the laser output power that are generated either during the lasing process or during the process of coupling into single-mode fibre. Intensity noise can be characterized in terms of relative intensity noise power spectral density (*RIN*) (i.e. the noise power in a 1 Hz bandwidth relative to the mean optical power) where

$$RIN[\text{dB Hz}^{-1}] = 10\log\left(\frac{\overline{\delta P^2}}{\overline{P}^2}\right). \tag{6.16}$$

Here, $\overline{\delta P^2}$ is the mean-square spectral density of power fluctuations and $\overline{P}$ is the mean laser power.

In single-mode channelled-substrate planar LDs, the intensity noise has been measured as −121 dB Hz^{-1} at 1 kHz which falls off with frequency at a rate of about 10 dB per decade [20]. In diode-pumped Nd:YAG ring lasers, the intensity noise was measured to be less than −110 dB Hz^{-1} between 100 Hz and 50 kHz [19]. The relaxation oscillation noise peak exhibited by these lasers can be suppressed using electronic feedback to the pump diode [21]. Suppression of the very low frequency intensity noise of these lasers has also been demonstrated using an external modulator with active feedback, which suppressed the intensity noise by 40 dB at 1 Hz from −65 dB Hz^{-1} to −105 dB Hz^{-1} [22]. In single-sensor Mach–Zender-based systems where both the outputs of the interferometer are available, a system of balanced detection can also be used to suppress intensity noise by common-mode rejection. Suppression by up to 40 dB has been demonstrated [23,24]; however, in most multiplexed systems, only one interferometer output is available preventing the implementation of this technique.

Laser frequency noise. It was shown in equation (6.2) that modulation of the laser frequency is converted to modulation of the total phase of a light wave propagating in a fibre. In an interferometer, frequency noise of the laser output is thus converted into intensity noise on the photodiode by the interferometer and this will degrade the phase resolution. Frequency noise of the laser source can be expressed in terms of the phase noise generated by an unbalanced interferometer. The measurement is usually normalized to an interferometer fibre imbalance of 1 m. For a Mach–Zender interferometer, the phase noise spectral density measured between the two output arms, $\delta\phi$, as a function of the frequency noise spectral density, $\delta\nu$, of the input signal is given by

$$\delta\phi = \frac{2\pi nd}{c}\delta\nu \tag{6.17}$$

where d is the fibre path imbalance and c is the light velocity in a vacuum. From equation (6.17), it can be seen that the magnitude of the phase difference is linearly proportional to the imbalance, d, in the interferometer. For the measurement, the path imbalance is set such that the laser-frequency-induced intensity noise is significantly larger than the laser intensity noise and any other noise sources present in the system, but not so large as to prevent linear demodulation of the phase. The phase noise spectral densities, in units of μrad Hz$^{-1/2}$ on the right-hand axis and frequency noise in Hz Hz$^{-1/2}$ on the left-hand axis, of four commercially available laser sources are shown in Figure 6.4.

The phase noise of the single-mode planar LD has been measured as 400 μrad Hz$^{-1/2}$ at 1 kHz [25]. A noise suppression

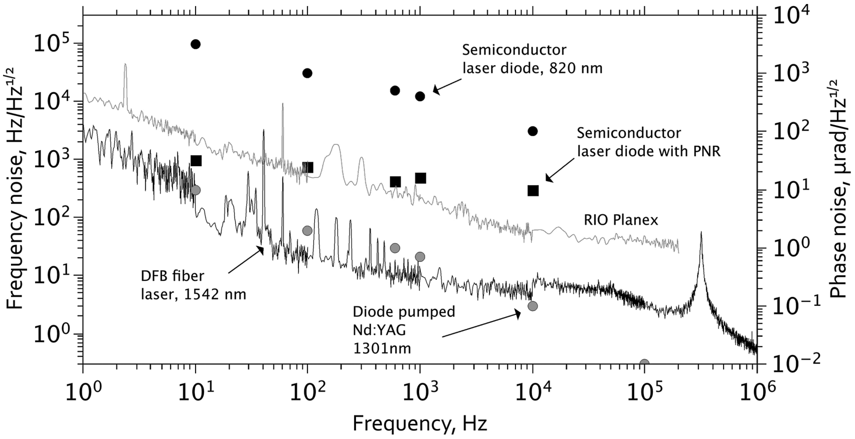


FIGURE 6.4 Laser phase noise for four commercially available lasers.

technique based on active feedback (utilizing the frequency modulation property of the LD described earlier) has also been demonstrated to suppress the phase noise by up to 40 dB over the frequency range 1 Hz to 10 kHz, also shown in Figure 6.4 [26]. It should be noted that this measurement was conducted on a Hitachi HLP 1400 diode laser at 820 nm; however, measurements on a DFB LD at 1310 nm have demonstrated similar phase noise levels. The frequency-induced phase noise of the diode-pumped Nd:YAG laser has been measured as 0.7 µrad $\text{Hz}^{-1/2}$ at 1 kHz [19]. Fibre lasers have also been characterized for use in interferometric sensors. A grating-stabilized semiconductor LD (RIO Planex™) generates a phase noise of 7 µrad $\text{Hz}^{-1/2}$ at 1 kHz [27]. Another type of fibre laser is the single-mode DFB Er^{3+}-doped fibre laser. The laser consists of a high Er^{3+} concentration doped fibre, with an in-fibre grating structure formed in the core to provide the resonant cavity. Pumped with either a 980 or 1480 nm source, the laser typically produces a few 100 µW of power, which can be amplified in a master-oscillator power amplifier configuration to a few mW. The phase noise of a DFB fibre laser source pumped with 100 mW at 1480 nm was measured in our laboratories to be 0.3 µrad $\text{Hz}^{-1/2}$ at 1 kHz. This type of laser is particularly promising due to its simplicity and small size and may prove to be an alternative to the LD.

It should be noted that, in the case of the Sagnac interferometer, the effective path imbalance is zero, and hence, the frequency noise property of the laser is less important. These sensors can be interrogated with broadband sources such as super luminescent diodes or super fluorescent fibre sources. However, it has been shown that interference between the primary waves and the Rayleigh backscatter waves generates excess noise through the conversion of source frequency noise to intensity noise [14].

6.3.5 Summary of Characteristics of Lasers Used in Interferometric Fibre Optic Sensors

In many applications, the performance of the laser source will determine the overall system performance and an appropriate choice of laser source is important for an optimized system design. A table summarizing the characteristics of each laser is shown in Table 6.2.

6.3.6 Components

Here, we describe the main optical components that make up a typical system.

Optical receivers. The optical receiver is required to convert the optical signal into an electrical voltage with the addition of minimal noise. In most cases, the PIN photodiode (based on InGaAs for operation at 1310 or 1550 nm) combined with a transimpedance amplifier (i.e. a PINFET) will give the best combination of bandwidth and responsivity; however, in some applications where light levels are low, the avalanche photodiode may give improved sensitivity.

TABLE 6.2

Characteristics of Laser Sources Used in Interferometric Fibre Sensors

Laser and Supplier	Wavelength (nm)	*RIN* at 1 kHz (dB Hz^{-1})	Phase Noise at 1 kHz (µrad $\text{Hz}^{-1/2}$ m^{-1})	Output Power (mW)
Single-mode LD	820/1310/1550	−121	400(16)[a]	>20
RIO Planex	1550	−141**	7**	10
Diode-pumped Nd:YAG ring laser[c]	1310	−118	0.7	>100
DFB fibre laser, NKT[d]	1550	−120[b]	0.3[b]	~10

[a] Figure in brackets is with noise suppression.

[b] Devices with lower noise are also available.

[c] Supplied by Lightwave Electronics, US.

[d] Supplied by NKT, Denmark.

Directional couplers. All of the interferometers shown in Figure 6.2 make use of fibre-directional couplers. These devices, which are an in-fibre analogue of the optical beam-splitter, are used to route signals through an optical fibre network. Their role in telecommunications applications has meant that, during the last decade, their cost has reduced significantly and their performance improved. Most commercially available couplers are based on the fused-fibre approach and provide a means of splitting light from one into two or more fibres, with a splitting ratio determined by the interaction length in the fused-fibre region. Commercially available couplers provide splitting ratios between 50:50 and 99:1, with an excess loss of less than 0.1 dB. The fused region can also be enclosed in a housing less than 3 mm in diameter and 25 mm in length.

Fibre Bragg gratings (FBG). These devices are formed by inscribing a periodic perturbation of the refractive index in the core of the optical fibre, usually with a ultra-violet laser (see Ref. [49]). This acts as a one-dimensional diffraction gratings and can be used as a wavelength-selective mirror. In optical hydrophone arrays, the FBG can be used as a mirror to implement in-line Michelson interferometers or as a means of separating wavelengths in a wavelength division multiplexed system. They can also be used to form high-finesse optical cavities and distributed feedback fibre lasers described later.

Acousto-optic modulators or Bragg cells. Acousto-optic modulators (AOMs) can be used both as a switch and a means of generating an optical frequency shift. The typical frequency shift is in the range 20–100 MHz. When used as a switch, the switching speed is a function of the material, the piezoelectric element bandwidth and the optical beam diameter—rise times of around 50 ns can be achieved. AOMs are now commercially available as fibre pig-tailed devices with insertion losses of around 5 dB.

Phase modulators. Phase modulators have a wide range of applications in hydrophone systems, for instance, as the control element in active homodyne systems, and in many of the remote interrogation schemes described later. They are normally based on two approaches, namely, piezoelectric fibre stretchers or integrated optic modulators (IOMs). In the first case, a length of fibre is wrapped around a piezoelectric element, usually a lead zirconate titanate (PZT) cylinder. A signal applied to the PZT causes a strain to be transferred to the fibre. In an IOM, the phase of an optical signal is modulated within a planar waveguide (typically made of lithium niobate or gallium arsenide) by the electro-optic effect. IOMs are pigtailed in single-mode fibre and exhibit an insertion loss ~2 dB. For a high-speed device, they require a defined input state of polarization in a polarization-preserving fibre but can modulate at speeds in excess of 1 GHz.

Amplitude modulators. The amplitude modulator provides a means of controlling the intensity of the light travelling in the fibre. In multiplexed sensor arrays, they can be used to generate the optical pulses used to interrogate the array. Although there are many designs of amplitude modulator, an efficient technique uses the electro-optic effect to modulate the phase of the light in one arm of an interferometer, thus modulating the output intensity of the interferometer. These are usually referred to as electro-optic switches. Use of the electro-optic effect allows high-speed modulation, and pulse widths less than 50 ns are achievable. An important specification of this device is the extinction ratio which characterizes the on-to-off output intensity of the switch. In multiplexed arrays, a low extinction ratio will cause light to be continuously injected into the array, which causes crosstalk between sensors and can be particularly severe if a highly coherent laser source is used. Most commercially available devices achieve ~30 dB when their operating point is actively stabilized against temperature drifts; however, by combining devices in series this can be significantly increased. For most applications, an extinction ratio of greater than 50 dB is adequate. Also, the polarization dependence of the electro-optic coefficient utilized in these devices necessitates the use of polarization-maintaining fibre on the input to these devices to ensure correct alignment of the input polarization state.

Semiconductor optical amplifiers – This device amplifies optical signals using a semiconductor junction. Polarization-independent devices are available with output saturation powers around 10 dBm and small signals gain up to 30 dB. Polarization-dependent devices offer higher saturation output powers. These devices can also be used at fast, high-extinction ratio switches. In their off state, their insertion loss can be as less than 60 dB which increases to 0 dB in their on state. This is far greater than that obtained from a single-stage IOM and also does not require stabilization of the operating point. Rise times of less than 5 ns can also be achieved enabling very fast switching.

Wavelength division multiplexing (WDM) devices – Devices that provide selective coupling between different wavelengths are referred to a wavelength division multiplexers and come in many different forms. The majority of these devices have been developed for telecommunications systems and operate in the 1550 nm transmission window (known as the C-band). Fused fibre couplers can be designed to combine different wavelengths onto a single fibre and are often used to combine pump wavelengths (e.g. 980 nm and 1480 nm) with signal wavelengths (e.g. 1550 nm). Thin-film filter devices are also widely available to perform similar functions to fused devices as well as for much higher finesse wavelength selectivity required in dense wavelength division multiplexing (DWDM). Thin-film devices are capable of separating wavelengths spaced by 50 GHz (0.4 nm) and are designed to operate with wavelengths positioned on the international telecommunications grid. High finesse tunable Fabry-Perot filters are also available to provide highly selective and tunable wavelength filtering. For handling large numbers of closely spaced wavelengths, a device known as an arrayed waveguide gratings can be used to separate wavelengths over the entire S, C and L band (covering the wavelength range from 1460 nm to 1625 nm) with a wavelength spacing of 50 GHz (i.e. greater than 100 channels). This level of functionality greatly exceeds that typically required in fibre optic sensors systems.

6.4 Acoustic Interactions

6.4.1 The Basic Transduction Mechanism

A practical optical fibre hydrophone design will combine an efficient pressure to phase transduction mechanism with a practical and simple design that meets the requirements listed in Section 6.2. This section considers various techniques for increasing hydrophone responsivity (defined as the optical phase shift per unit pressure) compared to that of bare fibre. We also discuss the issues involved in sensor design and present a range of practical hydrophone designs. First, the basic expression for the phase responsivity of an optical fibre is derived.

Consider a single-mode optical fibre of length L_s exposed to a static pressure, P, and acted on by an acoustic signal of amplitude, ΔP. We will assume that ΔP has no spatial dependency and is slowly varying such that the fibre is under quasi-static conditions. We also assume that the strain response of the fibre is linear (which will be true for strains up to ~1%). It was shown in equation (6.2) that the fractional phase change is governed by a length change effect and a refractive index (the stress-optic effect) term such that

$$\frac{\Delta\phi}{\phi} = \varepsilon_z + \frac{\Delta n}{n} \tag{6.18}$$

Here, we have neglected a third term related to the change in propagation constant due to the change in core diameter which can be shown to be small [70]. The second term is computed from the optical indicatrix,

$$\Delta\left(\frac{1}{n^2}\right)_{ij} = \sum_{kl} p_{ijkl}\varepsilon_{kl} \tag{6.19}$$

where ε_{kl} are the strain components (in Cartesian coordinates) and p_{ijkl} are the photoelastic components (the Pockel's coefficients). Fibres are typically fabricated from isotropic glass materials, so there are only two independent Pockel's coefficients (p_{11} and p_{12}). Provided that ΔP is uniform along the fibre length, there are no shear strains within the fibre and only three strain components are present [29], namely, $\varepsilon_z = \Delta l/l$ and $\varepsilon_\theta = \varepsilon_r = \Delta r/r$. It has been shown that the refractive index term for a fibre under uniform pressure is given by [30]

$$\Delta n = -\frac{n^3}{2}[\varepsilon_r(p_{11} - p_{12}) - \varepsilon_z p_{12}]. \tag{6.20}$$

Thus, the normalized responsivity, $\overline{R_P}$, *for a* fibre under pressure can be expressed in terms of the strain components in the fibre as

$$\overline{R_P} = \frac{\Delta\phi}{\phi\Delta P} = \frac{1}{\Delta P}\left[\varepsilon_z - \frac{n^2}{2}\varepsilon_r(p_{11} + p_{12}) + \varepsilon_z p_{12})\right]. \tag{6.21}$$

The acoustic responsivity of optical hydrophones is commonly expressed in the literature in one of two ways: (i) the responsivity R_P, given by $\Delta P/\Delta\phi$ expressed in rads Pa^{-1} and used with the interferometric phase resolution to determine the acoustic resolution of the sensor; or (ii) the normalized responsivity $\overline{R_P}$ defined earlier. This is a figure of merit independent of wavelength and fibre length and can be used to compare the transduction efficiency of various hydrophone designs directly. In the remainder of this section, we will express hydrophone acoustic responsivity in terms of the normalized responsivity and use a unit in decibels relative to 1 μPa^{-1}, where

$$\overline{R_P}[\text{dB re } \mu\text{Pa}^{-1}] = 20\log_{10}(\overline{R_P}/1\mu\text{Pa}^{-1}). \tag{6.22}$$

The responsivity of a bare fibre at low frequencies can be determined by calculating the strain in a hydrostatically loaded homogeneous fibre. In this case, it is straightforward to show that

$$\varepsilon_z = \varepsilon_r = -\Delta P\frac{2\sigma_g - 1}{E_g} \tag{6.23}$$

where E_g and σ_g are the Young's modulus and Poisson's ratio of the glass, respectively. Substituting this into equation (6.21) gives $\overline{R_P} = -351$ dB re μPa^{-1} where we have taken E_g = 72 GPa, $\sigma_g = 0.17$, $p_{11} = 0.121$, $p12 = 0.270$ and $n = 1.458$.

6.4.2 Coated Fibres

The previous analysis assumes an uncoated fibre but, in practice, all optical fibres make use of a plastic coating (normally in two layers). Fairly early in the development of the optical fibre hydrophone, it became clear that the pressure sensitivity of a length of coated fibre was higher than that of an uncoated fibre. It was shown by Budiansky et al. [30] that this would be expected when using compressible coatings with a diameter several times that of the fibre, and this analysis was extended by Hughes and Jarsynski to multi-layered coatings [32].

To calculate the responsivity of a coated fibre, the fibre is treated as a multi-layer cylinder with the glass forming the central core (the very small differences between the elastic properties of the core and cladding glass are normally neglected). The analysis calculates the stresses and strains in each layer of the structure using the Lamé solutions, deriving a set of simultaneous equations that can be solved assuming certain boundary conditions. Two boundary conditions have been used at the fibre ends. In the hydrostatic model, the pressure, ΔP, is assumed to act equally over the end surface of the fibre, so that the axial stress at the ends of the fibre is $-\Delta P$. In the radial model, it is assumed that no pressure acts on the ends of the fibre, so that the axial stress at the ends of the fibre is zero. For typical fibres, the greatest contribution to the responsivity is produced by the fibre length change (the first term in equation (6.18)). In general, the contribution to the responsivity due to the refractive index change (the second term in equation (6.18)) is of opposite sign and, hence, reduces the overall responsivity slightly.

The elastic properties of the coating materials are characterized in terms of the Young's modulus, E_c, and the Poisson's ratio, σ_c. Alternatively, the properties can be expressed in terms of E_c and the bulk modulus, K_c, where $K_c = E_c/3(1-2\sigma_c)$. Lagakos et al. [33] have carried out a detailed survey of suitable coating materials. It was found that for fibres with coatings of thickness less than 1 mm, the responsivity is a function of both E_c and K_c, while for 'thick' coatings, the responsivity is simply inversely proportional to K_c. The 'thick' coating limit is also reached faster for high E_c materials, as shown in Figure 6.5a.

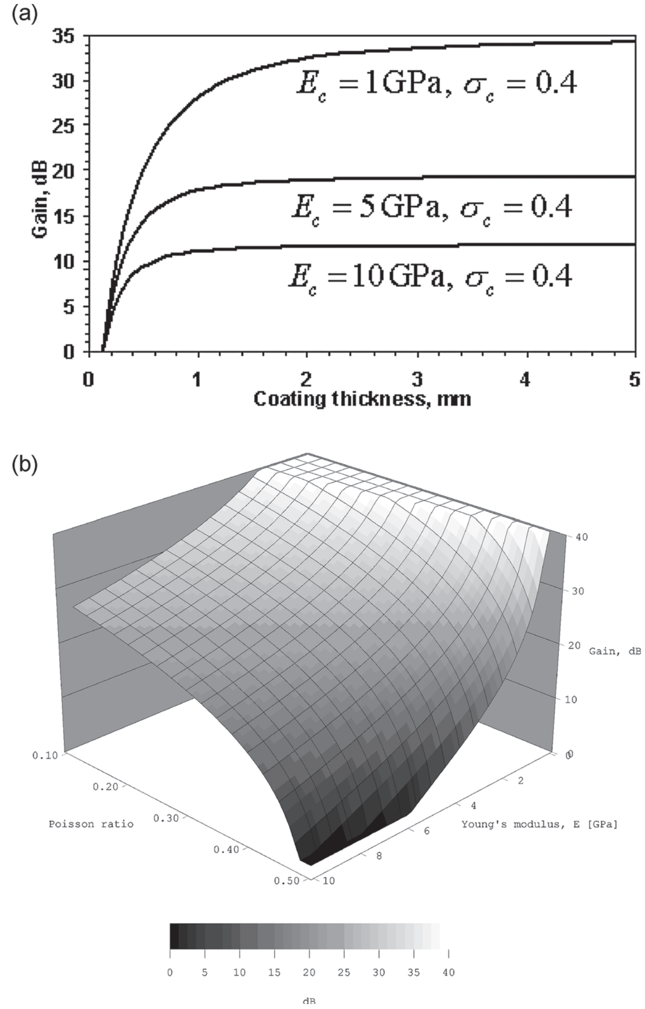


FIGURE 6.5 (a) Responsivity enhancement *versus* coating thickness and (b) responsivity enhancement *versus* E_c and σ_c for a thick coated fibre.

Here, it is shown that the thick coating limit is reached faster for the material with highest E_c, but that in the thick coating region, the responsivity is 22 dB higher for a material with an E_c of 1 GPa than for one with E_c of 10 GPa. Assuming a ‘thick’ coating is used, the highest responsivity is achieved with a low E_c and low K_c and, hence, a low σ_c (i.e. a coating with high compressibility). Such materials include Teflon™ and a range of epoxy resins and polyurethanes (e.g. Uralite™) [50]. By contrast, materials such as silicone rubber have a low E_c and high K_c (due to their high σ_c) and so produce a lower responsivity enhancement. It was shown in [34] that use of a thick (6 mm) coating of a material with optimum properties, such as Teflon™, can produce an increase in responsivity of around 30 dB compared to bare fibre. More recent work has demonstrated higher increases in responsivity greater than 50 dB can be obtained with air-included polymer coatings [100]. These air inclusions have the effect of further reducing the bulk modulus of the material.

Figure 6.5b shows how the responsivity, relative to a bare fibre, varies as a function of E_c and σ_c. This has been calculated using a 2D plain strain model for a 5 mm diameter, single-layer-coated fibre under hydrostatic boundary conditions. The properties of the fibre are the same as that used previously. A useful approximation for the responsivity of a thick-coated fibre, in terms of the bulk modulus of the coating, K_c, can be obtained by setting the radial strain in the fibre to zero such that

$$\overline{R_P} = \frac{1}{3K_c}\left(1-\frac{n^2 p_{12}}{2}\right). \tag{6.24}$$

It has been shown that this formula gives good agreement with experiment for coating diameters of around 1 mm or greater and for a wide range of coating materials [38].

Although fibres coated with compliant coatings offer an effective means of increasing the phase response of the optical fibre to static pressure, the resulting structure is soft. For optical fibres longer than a few centimetres, structural resonances occur at frequencies less than 100 Hz severely limiting the operating bandwidth of this structure. More recently, it has been shown how optical fibres respond to acoustically driven fluid motion [101,102] which can produce phase shifts of a similar magnitude to the pressure-induced phase shift. These additional phase shifts will be further amplified by flexural resonances of the structure. These effects likely explain why coated optical fibres exhibit unpredictable frequency-dependent response to pressure in underwater applications. A practical hydrophone design must be mechanically stable and avoid mechanical resonances occurring within the operating bandwidth. This bending motion observed in coated fibres and the associated modal resonances can be suppressed by winding the optical fibre in a coil, and this design forms the basis of the mandrel hydrophone.

6.4.3 Mandrel Hydrophones

To achieve the required hydrophone responsivity for most sonar applications, a larger amplification is required than that obtained from a ‘thick’-coated fibre, along with a broad operating bandwidth. Thus, the fibre must be wrapped in a coil to maintain a sensor size smaller than the acoustic wavelength. One design, known as the mandrel hydrophone, consists of a coil of fibre wound onto a tube or mandrel and encapsulated with a compliant material. A cross section of a generalized mandrel hydrophone is shown in Figure 6.6. This consists of a central cavity of radius r_1, with an internal pressure ΔP_i, a mandrel with an outer radius of r_2, a fibre layer (which is normally a fibre/encapsulant composite) with outer radius r_3 and an outer protective layer of encapsulant with a radius, r_4. The mandrel is typically a tube of metal or plastic and may be solid (i.e. $r_1 = 0$) or may be air-backed (i.e. the central cavity is air-filled and sealed from the external pressure).

In one configuration of the mandrel hydrophone, shown in Figure 6.7, the mandrel is actually a layer of the same encapsulant used in the outer layer. In this case, the fibre layer is effectively embedded within a thick tube of encapsulant. Lagakos and Bucaro [34] treated such a hydrophone as a straight coated fibre having a thick coating with a diameter of $r_4 - r_2$. This approach gave results very close to those

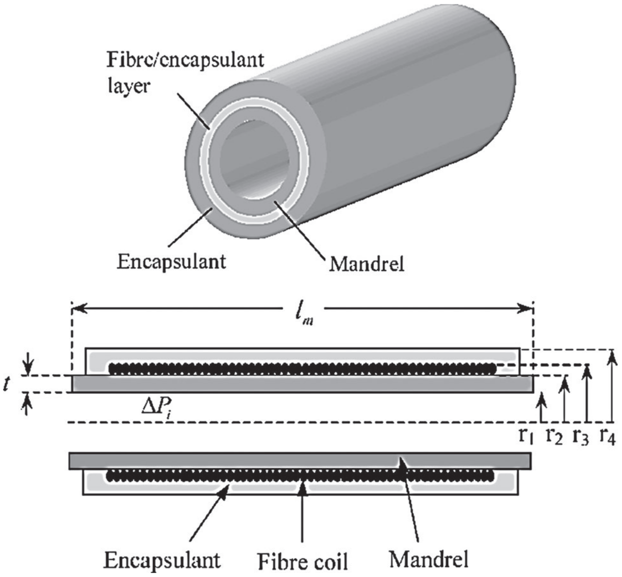


FIGURE 6.6 Mandrel hydrophone cross section.

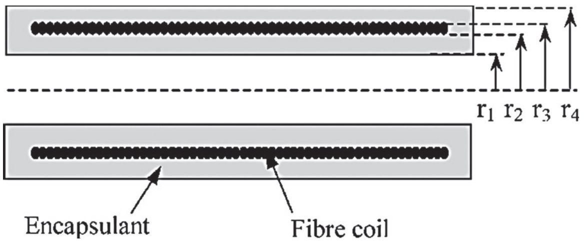


FIGURE 6.7 Embedded hydrophone cross section.

obtained from experiment. For a hydrophone fabricated using Uralite™ polyurethane, a normalized responsivity of –320 dB re μPa^{-1} was obtained. An alternative approach (that can be used to model both the mandrel designs shown in Figures 6.6 and 6.7), reported by Nash and Keen [35], treats the mandrel hydrophone structure as a multi-layer cylinder and derives the strains in the cylinder layers using the same analysis as previously described for a coated fibre. In this case, it is assumed that the fibre is constrained by the dimensional changes in the mandrel, so that the axial strain in the fibre is equivalent to the hoop strain in the cylinder approximately at radius r_2. The glass within the fibre layer must be taken into account, as this has a significant stiffening effect on the hydrophone. The fibre layer forms a glass/encapsulant matrix and its elastic properties can be approximated by taking a volume average of the glass and encapsulant. As with a coated fibre, this model gives accurate results for a long, thin cylinder, so it is more accurate for hydrophone designs in which the length of the coil is significantly greater than its diameter. It has also been shown to achieve good agreement with theory. A hydrophone encapsulated in Araldite™ epoxy resin was predicted to achieve a normalized responsivity of –324.5 dB re μPa^{-1} and results within 2 dB of this prediction were obtained experimentally. In general, for the embedded coil hydrophone, the highest responsivity is obtained with a rigid interior and a low bulk modulus encapsulant.

This model can also be directly applied to the case of an air-backed mandrel by setting the internal signal pressure, ΔP_i, to zero (this approximation can be used because air can be considered infinitely compressible compared to water). An air-backed mandrel offers the potential for significantly higher responsivity than other designs, because the presence of air increases the overall compressibility of the cylinder. For instance, with an air-backed mandrel hydrophone design, a normalized responsivity of –310.5 dB re μPa^{-1} was obtained [35]. Similar results have been obtained elsewhere; for instance, Knudsen et al. [36] obtained normalized responsivities in the range –307 to –315 dB re μPa^{-1} for similar air-backed mandrel designs. The highest normalized responsivity reported to date was by Wang et al. [37], who achieved –298 dB re μPa^{-1} using an air-backed polycarbonate mandrel with a single fibre layer. The mandrel design has, thus, been demonstrated to provide up to 53 dB in amplification of the pressure responsivity compared to a bare fibre. Most air-backed mandrel hydrophone designs use up to 100 m of optical fibre and thus responsivities, R_P, around 0 dB re rad Pa^{-1} (1 rad Pa^{-1}) are easily achievable.

In air-backed mandrel hydrophones, with a moderate number of fibre layers, it is reasonable to assume that the axial strain in the fibre is significantly larger than the radial strain. Thus, by ignoring the radial strain in the fibre, a simplified formula for the hydrophone responsivity can be derived. It was shown by Knudsen [38] that for an infinitely long, two-layer, air-backed mandrel hydrophone,

$$\overline{R_P(r)} = \frac{0.78}{E_{\text{rmf}}}\left\{-\frac{C_1}{r^2}(1+\sigma_{\text{f}}) + 2C_2(1-\sigma_{\text{f}})\right\} \tag{6.25}$$

where the constants C_1 and C_2 are given by

$$C_1 = -r_3^2 - \left[r_3^4 \frac{E_{\text{m}}(r_2^2 - r_1^2)(1+\nu_{\text{f}}) + E_{\text{f}}(r_2^2(1-\sigma_{\text{m}}) + r_1^1(1+\sigma_{\text{m}}))}{E_{\text{m}}(r_2^2 - r_1^2)\{r_3^2(1+\sigma_{\text{f}}) - r_2^2(1-\sigma_{\text{m}})\} - E_{\text{f}}(r_3^2 - r_2^2)\{r_2^2(1-\sigma_{\text{m}}) + r_1^2(1+\sigma_{\text{m}})\}} \right]$$

$$C_2 = r_3^2 \frac{E_{\text{m}}(r_2^2 - r_1^2)(1-\nu_{\text{f}}) + E_{\text{f}}(r_2^2(1-\sigma_{\text{m}}) + r_1^1(1+\sigma_{\text{m}}))}{2E_{\text{m}}(r_2^2 - r_1^2)\{r_3^2(1+\sigma_{\text{f}}) - r_2^2(1-\sigma_{\text{m}})\} - E_{\text{f}}(r_3^2 - r_2^2)\{r_2^2(1-\sigma_{\text{m}}) + r_1^2(1+\sigma_{\text{m}})\}}.$$

This equation is based on a two-layer model, where the outer layer of encapsulant is included in the fibre layer, so that $r_3 = r_4$ in Figure 6.6. Here E_m, E_f, σ_m and σ_f are the Young's modulus and Poisson's ratio of the mandrel (between r_1 and r_2) and the fibre/encapsulant composite (between r_2 and r_3), respectively. The effective properties for the composite layer are approximated using a simple law of mixtures based on the respective volume fractions of glass and encapsulant. For a hydrophone with multiple fibre layers, equation (6.26) is calculated at a radius, r, corresponding to the centre of the fibre layer. This equation is expected to be accurate when the stiffness of the mandrel is much higher than the stiffness of the fibre layer.

In general, air-backed designs will offer the highest normalized responsivity but the poorest resistance to static pressure. For most practical designs, the use of an air-backing provides an increase of the normalized responsivity by up to ~20 dB compared to a solid mandrel, although greater increases can be achieved for very shallow water designs.

The analytical hydrophone models described here are based on long, thin cylinder designs, and they are less accurate when the hydrophone diameter is of the same order as its length; however, approximate correction factors can be introduced for these geometries. The models are also unsuitable for non-symmetric or other, more complex, hydrophone designs. In this case, a finite element approach is more suitable. Such an approach, described in [39], has been shown to give results equivalent to the analytical models for simple hydrophone designs in hydrostatic conditions.

6.4.4 Hydrophone Responsivity with Depth

The limit on maximum hydrophone responsivity is imposed by the requirement for maximum operating and survival depth. The requirement for maximum operating depth is that the strain imposed in the fibre due to the ambient pressure does not exceed the linear limit for fibre (typically ~1%). The requirement for maximum survival depth is (i) that the fibre strain due to the ambient pressure must not exceed its breaking strain (typically ~4%) and (ii) that the hydrophone structure does not collapse or buckle.

Assuming that the hydrophone structure is rigid, then we can calculate the maximum survival depth of the hydrophone from the requirement that the fibre strain does not exceed 4%. If the fibre is under axial strain only, then using equation (6.21) and taking $\varepsilon_r = \varepsilon_z = \varepsilon$, it can be shown that the responsivity reduces to [70]

$$\overline{R_P} \approx \frac{0.7}{\Delta P}\varepsilon. \tag{6.26}$$

Setting ε equal to 0.04 (i.e. 4%) and ΔP to the required survival pressure, the maximum hydrophone responsivity can be approximated. The maximum operational pressure will be a factor of four less (corresponding to 1% strain). The maximum hydrophone responsivity *versus* survival depth is plotted in Figure 6.8, where we have taken 10 m ≡ 10^5 Pa.

It can be seen that if the hydrophone is required to survive to 1000 m (a typical requirement), the maximum allowed responsivity would be −291 dB re μPa^{-1}. For operation down to 500 m, the survival depth will be 2000 m (i.e. four times 500 m) and the responsivity must be no greater than −297 dB re μPa^{-1}. Thus, the specified maximum survival and operating depth will, therefore, set an upper limit on the normalized responsivity.

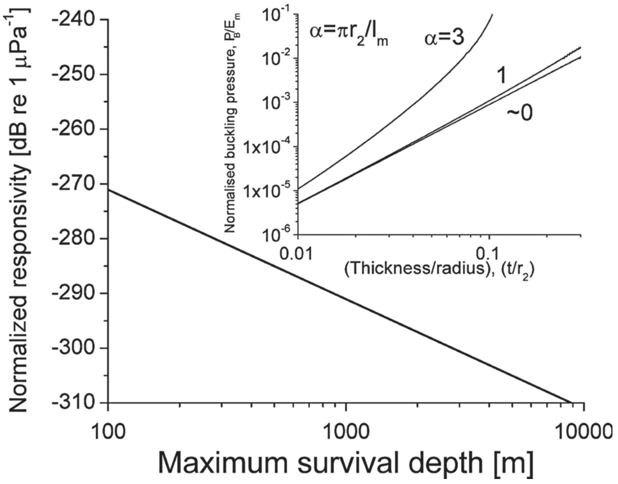


FIGURE 6.8 Maximum hydrophone responsivity *versus* survival depth.

Preventing the hydrophone structure from buckling becomes difficult at great depths and when air-backing is incorporated into the hydrophone. The buckling pressure of a hydrophone can be estimated by considering the buckling pressure of a thin-walled, simply supported cylinder under hydrostatic pressure, which represents the hydrophone mandrel. It was shown by Farshad [38,57] that the lowest critical pressure for a buckling mode, P_B, is given by

$$P_\mathrm{B} = \frac{E_\mathrm{m}}{12(1-\sigma_\mathrm{m}^2)}\left(\frac{t}{r_2}\right)^3 \frac{((\pi r_2/l_\mathrm{m})^2+\gamma^2)^2}{\gamma^2} + E_\mathrm{m}\left(\frac{t}{r_2}\right)\frac{(\pi r_2/l_\mathrm{m})^4}{\gamma^2((\pi r_2/l_\mathrm{m})^2+\gamma^2)^2} \tag{6.27}$$

where l_m is the mandrel length, t (= $r_2 - r_1$) is the mandrel wall thickness and γ is the number of circumferential periods. This takes a value, corresponding to the lowest buckling pressure, given by the integer closest to

$$\gamma^2 = \frac{2^{1/2}\,3^{1/4}(1-\sigma_\mathrm{m}^2)^{1/4}}{(t/r_2)^{3/4}} - \left(\frac{\pi r_2}{l_\mathrm{m}}\right)^2. \tag{6.28}$$

Using equation (6.27), it can be shown that the buckling pressure is almost independent of the Poisson ratio of the mandrel and is approximately proportional to $(t/r_2)^2$. This is shown in the inset of Figure 6.8 for various values of the parameter, $\alpha = \pi r_2/l_\mathrm{m}$. Here, we have taken $\sigma_\mathrm{m} = 0.4$. For example, if E_m = 3GPa, l_m = 100 mm, r_2 = 8 mm, t = 1 mm, then α = 0.25, t/r_2 = 0.125 and the buckling pressure, P_B ~ 3MPa (or ~300 m depth).

A technique to alleviate the problem of buckling of the mandrel based on pressure compensation has been attempted on flexural-disc-based transducers [40,41]. This involves flooding the air cavity by incorporating a small hole in the hydrophone. The hole acts as an acoustic low-pass filter such that the ambient pressure on either side of the coil will be balanced but the acoustic signal will only impinge on the outer surface of the coil. The effectiveness of the design depends on the flooded cavity design and hole size. The sensor will also exhibit a low-frequency roll-off in responsivity dependent on the hole size.

It is interesting to deduce the dynamic range of an interferometric sensor from the elastic limit of glass. Assuming the strain sensitivity of the interferometer is 10^{-14} in 100 m of fibre and the elastic limit is 0.01, the available dynamic range is around 12 orders of magnitude or 240 dB!

6.4.5 Hydrophone Response at Higher Frequencies and Directionality

The responsivity of the sensor will depend on the dynamic response of the sensor structure. All structures exhibit mechanical resonant behaviour that is usually a complicated function of the mechanical properties of the materials and the sensor dimensions. To ensure the hydrophone exhibits a flat frequency response and a linear phase response, the hydrophone must be designed such that the frequency of operation is far below any mechanical resonances. Careful design of the sensor can also allow damping to be incorporated to reduce the effects of any resonances present.

Assuming the structure is free from mechanical resonances, it is possible to deduce the optimum dimensions of the mandrel hydrophone to obtain a near omni-directional response. This has been carried out by Knusden [38] for the simplified case where quasi-static bending effects are ignored and the effective responsivity per fibre loop is assumed to be constant across the mandrel. It was shown that the ratio of hydrophone radius to length should be approximately 0.4. Thus, if an omni-directional hydrophone response is required, the sensor volume must be decreased as the acoustic frequency increases. Also, for hydrophones with fibre lengths greater than 100 m and frequencies above about 20 kHz (i.e. for normalized frequency greater than about 0.4 for the Michelson interferometer in Figure 6.3), the transit time of the optical signal in the coil will also reduce the responsivity, as described in Section 6.3.

Finite element modelling has been used to predict the responsivity of optical hydrophones [39], and although attempts have also been made to analytically determine mechanical response, the FE approach is expected to give higher accuracy particularly for complex designs.

6.4.6 Sensitivity to Other Effects

So far we have only discussed the issues concerned with maximizing the hydrophone responsivity to a pressure change associated with an acoustic wave. Since the propagation of an acoustic wave involves mass flow, the medium will also exhibit a particle velocity that is related to the pressure change through the *linear inviscid force equation* [42]. In practice, the hydrophone will be sufficiently large (and hence of high inertia) to prevent any motion induced by the particle velocity. However, for many platform-mounted sonar arrays, the hydrophones may be exposed to significant vibration and, hence, acceleration forces, due to the motion of the platform.

It is, thus, important to design sensors for platform-mounted sonar arrays that exhibit low responsivity to acceleration. If we consider a mandrel hydrophone, of the type shown in Figure 6.6, subject to an acceleration force acting along the central axis of the mandrel, asymmetry in the hydrophone design will cause deformation of the encapsulant and fibre coil and, hence, a net strain in the fibre coil. This has been shown to be the primary cause of high-acceleration sensitivity in air-backed mandrel-type hydrophones [43]. It was shown that mounting the sensor, such that excitation is symmetrical about the fibre coil, can significantly reduce this acceleration sensitivity.

6.4.7 Practical Hydrophone Designs

The mandrel hydrophone is a commonly used design in many sonar systems under development. In particular, it lends itself for use in both towed and seabed arrays where the sensor must be of small diameter and easy to incorporate into the array structure. Typically, a sensor may comprise up to 100 m of fibre that must be wound onto a mandrel less than 15 mm in diameter. Standard telecommunications fibre such as SMF 28™ would exhibit unacceptably high levels of macro-bending loss if wound on this diameter due to its relatively low numerical aperture ($NA \simeq 0.11$). Increasing the numerical aperture of the fibre ensures that the light is more tightly confined to the fibre core, and hence, the fibre will be more resistant to bending. It is straightforward to increase the fibre *NA* to around 0.20 without a significant increase in the intrinsic attenuation in the waveguide [44,45]. These fibres can be wound onto mandrels of less than 10 mm diameter without observing any significant excess loss associated with macro-bending. It was also mentioned earlier that the responsivity of the mandrel hydrophone is strongly dependent on the stiffness of the fibre/encapsulant layer. Standard telecommunications fibre has a cladding/coating diameter of 125/250 μm, but by using a reduced diameter fibre such as 80/170 μm, an increase in the responsivity of up to a factor of two can be achieved due to the reduction in the volume of glass (and, hence, the reduced stiffness of the fibre layer).

The pressure resolution achieved by the sensor depends on both the hydrophone responsivity and interferometric phase resolution (discussed in the next section). Hence, a reduction in the hydrophone responsivity can be compensated for by an increase in the interferometric phase resolution. This would allow, for instance, an increase in the maximum operational depth of the hydrophone.

Both the Mach–Zender and Michelson interferometer configurations require an insensitive reference coil in addition to a sensor coil. Some designs of mandrel hydrophone incorporate this coil within the sensor coil. The reference coil is usually shielded from the acoustic signal and wound on a stiff metal mandrel to minimize its acoustic sensitivity. A typical arrangement of a two-coil air-backed mandrel hydrophone is shown in Figure 6.9.

In addition to the mandrel design, a wide range of other hydrophone configurations have been demonstrated. These include planar hydrophones based on a flat fibre coil [46], acceleration sensitivity reduced flexural disc designs [40,47]

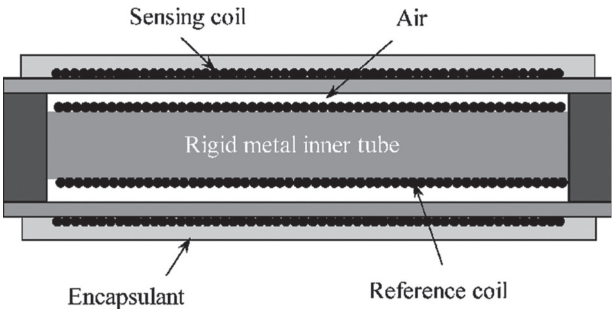


FIGURE 6.9 Typical mandrel hydrophone design.

and ellipsoidal flextensional hydrophones [48]. A brief list of reported hydrophone designs, together with their normalized responsivities, is given in Table 6.3. Although the list is not intended to be exhaustive, it gives an indication of the variety of hydrophone designs proposed.

Finally, Figure 6.10 shows examples of two hydrophone designs developed at the Defence Evaluation and Research Agency for towed and hull array sonar applications, respectively. These are both based on air-backed mandrel designs of the type shown in Figure 6.6.

6.5 Signal Processing

In a practical fibre optic interferometric sensor system, a passive interrogation technique is required to overcome the fading problem described in Section 6.2 (not to be confused with polarization-induced signal fading described later). Several techniques have been demonstrated that allow the modulating signal of interest to be extracted from the optical carrier signal [51–55] and are reviewed in [56]. The relative merits of each technique can be compared in terms of the phase resolution achieved and the ease of implementation. For application to sensor arrays, the technique must also be combined with a multiplexing technique. This section will describe two techniques to interrogate fibre interferometers remotely and present practical implementations of each. These two techniques have been the most widely used, to date, in practical hydrophone systems. A noise analysis will also be presented that will allow calculation of the phase resolution achievable with such interrogation schemes. Finally, a discussion on the dynamic range and the use of digital techniques for signal demodulation is given.

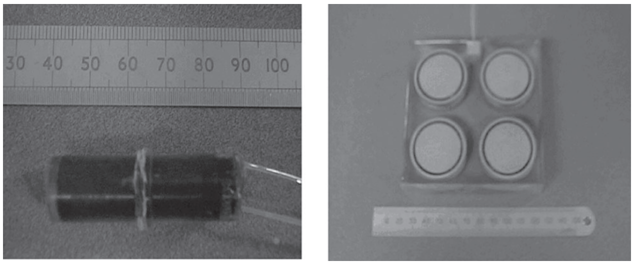


FIGURE 6.10 DERA hydrophone designs.

6.5.1 Passive Interrogation Schemes

The basic interferometric optical fibre sensor arrangement is shown in Figure 6.1. Optical modulators can be placed in the arms of the interferometer in positions labelled 1 and 2 that can either modulate the phase of the light (i.e. using a piezoelectric fibre stretcher) or impose a frequency shift (i.e. using an AOM) on the light. In the following analysis, we assume, for simplicity, that the optical losses in the signal and reference arms are zero and the coupling coefficients are 1/2, hence $\alpha_s = \alpha_r = 1$ and $\kappa_1 = \kappa_2 = 1/2$ in equation (6.6).

Homodyne phase-generated carrier (PGC). This scheme involves generation of a carrier frequency through high-frequency modulation of the optical phase and can be implemented in two ways: (i) incorporating a phase modulator in position 1 in Figure 6.1 allows a high-frequency phase modulation to be imposed on the optical signal in arm 1 of the interferometer; or (ii) modulating the optical frequency of the laser output and incorporating a small path imbalance in the interferometer, as shown in Figure 6.11, allow generation of a carrier through frequency discrimination [55]. This latter technique allows remote interrogation of the interferometer without the need for active components in the interferometer and its implementation with semiconductor LDs is now discussed.

The PGC interrogation scheme can be implemented by incorporating a small optical path imbalance in the interferometer and applying a modulation to the semiconductor LD. If a sinusoidal current modulation of angular frequency ω_{pgc} is added to the LD drive current, as shown in Figure 6.11, then the photocurrent in equation (6.7) becomes

$$i_{ph-pgc} = rP + rPV\cos[\phi_{pgc}\cos\omega_{pgc}t + \phi_s(t) + \phi_d(t)] \tag{6.29}$$

TABLE 6.3

Reported Hydrophone Performance

Design	Author	Norm Resp (dB re μPa^{-1})	Bandwidth	Ref.
Solid mandrel	Nash and Keen	–324.5	>1 kHz	[35]
Air-backed mandrel	Nash and Keen	–310.5	>1 kHz	[35]
Air-backed mandrel	Knudsen	–314	10 kHz	[36]
Flexible planar	Lagakos	–318	>2.5 kHz	[46]
Solid mandrel	Lagakos	–320	10 kHz	[34]
Air-backed mandrel	Wang	–298	5 kHz	[37]
Flexural disc	Cheng	–283	500 Hz	[40]
Air-backed mandrel	Cranch	–307.5	>6 kHz	

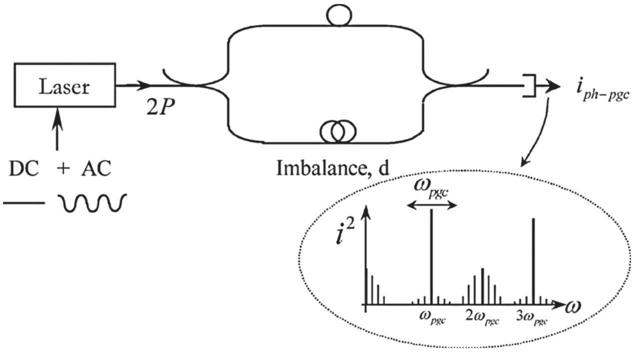


FIGURE 6.11 Laser frequency modulation implementation of homodyne PGC.

where $\phi_{\text{pgc}} = 2\pi nd\Delta v/c$, d is the path imbalance, c is the speed of light in a vacuum and Δv is the peak frequency excursion of the laser output (determined by the amplitude of the current modulation), and we have replaced the signal and reference arm phase with a time-varying signal phase and a slowly varying drift phase, $\phi_{\text{d}}(t)$, due to environmental effects. Expanding equation (6.29) in terms of its Bessel function components using the following relationships,

$$\cos(\phi_0 \sin(\omega_{\text{s}}t)) = J_0(\phi_0) + 2\sum_{k=1}^{\infty} J_{2k}(\phi_0)\cos(2k\omega_{\text{s}}t)$$

$$\sin(\phi_0 \sin(\omega_{\text{s}}t)) = 2\sum_{k=1}^{\infty} J_{2k+1}(\phi_0)\sin((2k+1)\omega_{\text{s}}t) \qquad (6.30)$$

and setting $\phi(t) = \phi_{\text{d}} + \phi_{\text{s}}$, yields

$$i_{\text{ph-pgc}} = rP + rPV\left[\left[J_0(\phi_{\text{pgc}}) + 2\sum_{k=1}^{\infty}(-1)^k J_{2k}(\phi_{\text{pgc}})\cos 2k\omega_{\text{pgc}}t\right] \times \cos\phi(t) - \left[2\sum_{k=0}^{\infty}(-1)^k J_{2k+1}(\phi_{\text{pgc}})\cos(2k+1)\omega_{\text{pgc}}t\right]\sin\phi(t)\right]. \qquad (6.31)$$

This shows that the photocurrent consists of a fundamental frequency component at ω_{pgc} (referred to as the carrier frequency) plus a series of harmonic components of the carrier frequency about which the signal of interest, $\phi(t)$, (= $\phi_{\text{s}}(t)$ + $\phi_{\text{d}}(t)$) appears as phase-modulation sidebands (shown in the inset of Figure 6.11). The amplitude of each harmonic is thus given by the respective Bessel function amplitude. This signal is then processed (or demodulated) to yield the sine and cosine of $\phi(t)$ by synchronous detection of the PGC components ω_{pgc} and $2\omega_{\text{pgc}}$, as shown in Figure 6.12a. By low-pass filtering, the resulting signals, the sine and cosine of the signal phase, are obtained:

$$rPVJ_1(\phi_{\text{pgc}} \sin\phi(t)) \qquad (6.32)$$

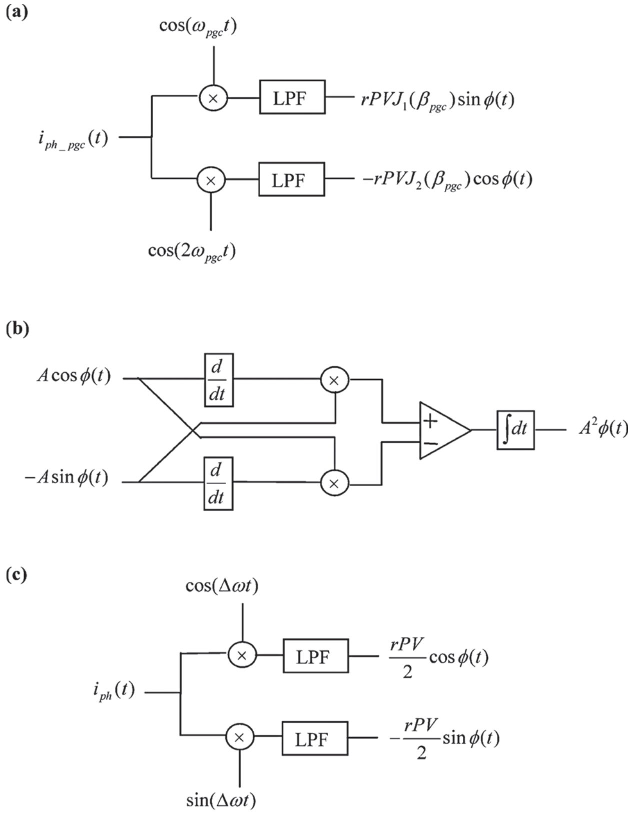


FIGURE 6.12 (a) Obtaining the sine and cosine components in the PGC scheme, (b) obtaining the sine and cosine components in the heterodyne scheme and (c) the differentiate and cross-multiply demodulator [55].

$$rPVJ_2(\phi_{\text{pgc}} \cos\phi(t)) \qquad (6.33)$$

$\phi(t)$ can then be obtained directly by setting the LD modulation current such that $J_1(\phi_{\text{pgc}}) = J_2(\phi_{\text{pgc}})$ (i.e. $\phi_{\text{pgc}} = 2.6$ or 6.1) and taking the arctangent of the ratio of equations (6.32) and (6.33). An alternative method uses a technique known as differentiate and cross-multiply and is shown schematically in Figure 6.12b. If the differential of equations (6.32) and (6.33) is taken and multiplied by the undifferentiated version of the other signal, then the difference between the two resulting signals yields

$$(rPV)^2 J_1(\phi_{\text{pgc}})J_2(\phi_{\text{pgc}})\dot{\phi}(t). \qquad (6.34)$$

Integration of this signal followed by normalization to remove the unwanted amplitude term yields the signal phase of interest, $\phi(t)$. It is worth noting that this technique allows the phase to be directly obtained with no ambiguity about 2π. Use of the arctangent method requires a fringe counting technique that monitors phase crossing at integer multiples of 2π in order to track phase excursions greater than 2π.

Heterodyne. In the heterodyne interrogation scheme, the frequency of the optical signal propagating in one arm of the interferometer is shifted relative to the other using the AOM. By placing an AOM in position 1 in Figure 6.1 and applying an rf signal to the AOM, a frequency shift equal to the frequency of the rf drive signal will be imposed on the diffracted beam. The photocurrent is then given by equation (6.7) where

$\Delta\omega$ is equal to the rf drive signal of the AOM. By placing a second AOM in position 2 in Figure 6.1, $\Delta\omega$ can be set to any value by varying the frequency of the rf signals driving each AOM. This technique yields a signal simpler in form than that obtained with homodyne PGC since it contains only a single-carrier frequency component at $\Delta\omega$. Extraction of the signal phase of interest, $\phi(t)$, can be achieved either with an electronic frequency discriminator and integrator or by mixing the photodiode signal with in-phase and quadrature versions of the unmodulated carrier frequency and low-pass filtering, as shown in Figure 6.12c. This yields

$$0.5r\ PV\cos\phi(t) \tag{6.35}$$

$$-0.5r\ PV\sin\phi(t) \tag{6.36}$$

which are the familiar sine and cosine of the signal phase. Extraction of $\phi(t)$ is then achieved using the arctangent function or the differentiate and cross-multiply techniques described earlier. It will be shown later how this technique can be used to interrogate multiplexed fibre interferometers remotely.

6.5.2 Noise Sources and Phase Resolution

Having derived the form of the received optical signal, the various sources of noise in the interferometric system can be included to calculate the carrier-to-noise ratio (CNR). This is then related to the minimum detectable phase or phase resolution by using narrow-band phase-modulation theory [58]. The following analysis is based on a heterodyne system although it can also be applied, with some minor modification, to a phase-generated carrier scheme.

Taking the ac part of the photocurrent in the case of a heterodyne system gives

$$i_{\text{ph-het}} = rPV\cos(\Delta\omega t + \phi_s(t) + \phi_d(t)) \tag{6.37}$$

Ignoring environmentally induced phase perturbations and assuming a sinusoidal signal excitation, $\phi_s(t)$ equal to ϕ_0 sin $\omega_s t$, this can be expressed as,

$$\begin{aligned} i_{\text{ph-het}} = rPV[&\cos(\Delta\omega t)\cos(\phi_0\sin(\omega_s t)) \\ &- \sin(\Delta\omega t)\sin(\phi_0\sin(\omega_s t))]. \end{aligned} \tag{6.38}$$

Using the equations (6.30) and the small-angle approximations, $J_0(\phi_0)\approx 1$, and $J_1(\phi_0)\approx\phi_0/2$ for $\phi_0\approx 1$, equation (6.38) can thus be expressed as

$$i_{\text{ph-het}} \approx rPV\left(\cos\Delta\omega t - \frac{\phi_0}{2}[\cos((\Delta\omega-\omega_s)t) - \cos((\Delta\omega+\omega_s)t)]\right) \tag{6.39}$$

The power ratio between the carrier and first modulation sideband is, therefore, given by

$$\text{CNR} \approx \left(\frac{2}{\phi_0}\right)^2. \tag{6.40}$$

Thus, to achieve a phase resolution of 1 μrad Hz$^{-1/2}$, a CNR of 126 dB re Hz$^{-1/2}$ is required (assuming a signal-to-noise ratio of 1.0 is required to resolve the phase). Using equations (6.37) and (6.40), the following expression can be derived that relates the sensor rms phase resolution, $\delta\phi$, to the total noise current, $\overline{i_n^2}$,

$$\frac{0.5(rPV)^2}{\overline{i_n^2}} \approx \left(\frac{\sqrt{2}}{\delta\phi}\right)^2 \tag{6.41}$$

where the factor of 0.5 converts peak to rms carrier power.

It now remains to determine $\overline{i_n^2}$. Fundamentally, $\overline{i_n^2}$ is limited by shot noise; however, in practice, other sources of noise will be present. These are (i) the detector amplifier thermal noise, (ii) intrinsic thermal noise in the fibre waveguide, (iii) rf oscillator noise, (iv) laser intensity noise, (v) laser frequency or phase noise and (vi) spontaneous emission noise from an optical pre-amplifier if present.

Expressions for the mean-square noise current due to shot noise and amplifier thermal noise are given by

$$\overline{i_{sh}^2} = 2e^2\eta P\Delta f / h\nu \tag{6.42}$$

$$\overline{i_{th}^2} = 4kT_e\Delta f / R_L \tag{6.43}$$

where e is the electron charge, k is Boltzmann's constant, $T_e = T + T_a$, T is the ambient temperature and T_a is effective noise temperature of the amplifier [59], R_L is the photodiode load resistance and Δf is the electrical bandwidth. If it is assumed that the carrier frequency is sufficiently high that the photodiode generation-recombination noise is negligible [60], then, in principle, the optical power can be increased such that $\overline{i_{sh}^2} = \overline{i_{th}^2}$ and shot-noise-limited detection is achieved. However, it has been shown that, in practice, it will be difficult to achieve resolutions approaching the shot noise limit due to the intrinsic thermal noise in the optical waveguide, which can be significant even for relatively short fibre lengths [61,62]. Thermodynamically driven fluctuations in the refractive index of optical waveguide impose a phase noise on the optical signal. Experimental measurement of this noise has shown that for a Mach–Zender interferometer with 100 m of fibre in each arm, the thermal-noise-limited phase resolution is about 0.8 μrad Hz$^{-1/2}$ at 2 kHz [62], limiting the CNR to 125 dB re Hz$^{-1/2}$, regardless of optical power.

In the heterodyne system, the AOMs must be driven by an rf signal, typically a few tens of MHz. This signal must be derived from a stable oscillator to ensure excess noise is not imposed by the AOM. For example, the rf oscillators must be chosen such that their relative noise power spectral density is less than −126 dB re Hz$^{-1/2}$ about the oscillator frequency if 1 μrad Hz$^{-1/2}$ resolution is required.

Intensity noise from the laser induces amplitude modulation of the optical signal, causing the RIN spectrum to appear as modulation sidebands around the carrier. The phase noise spectral density contribution due to RIN can be expressed as

$$\delta\phi_{\text{laser-RIN}} = \sqrt{RIN(f)}. \tag{6.44}$$

As discussed in Section 6.3, laser frequency noise will generate phase noise in an unbalanced interferometer. In one system, suppression of this phase noise was achieved using a reference interferometer. The phase noise was measured by the reference interferometer and subtracted from the sensor interferometer. This technique has been shown to suppress the laser phase noise by up to 30 dB at 50 Hz [64,65].

Noise from optical amplifiers. The final source of noise to be discussed is that produced by an optical pre-amplifier, which can be used to increase the optical power at the photodiode. In the following discussion, it is assumed that the pre-amplifier is placed immediately before the photodetector with a narrowband optical filter on its output as shown in Figure 6.13. The optical filter is assumed to have negligible insertion loss.

Amplification of light by stimulated emission will introduce a broadband incoherent noise due to spontaneous emission that, on detection, generates an excess noise current generally larger than the shot noise and detector thermal noise contribution. The noise generated arises from the following three effects: (i) beating of the spontaneous emission with itself, (ii) beating of the spontaneous emission with the signal and (iii) beating of the spontaneous emission with the signal shot noise; these have been studied extensively for optical communications networks. Expressions for these three effects in terms of the mean-square noise current generated are given by equations (6.45)–(6.47), respectively, and the modified shot noise current is given by equation (6.48) (see Ref. [66]),

$$\overline{i_{\mathrm{sp-sp}}} = (e\eta GF)^2 \Delta\nu_{\mathrm{opt}} \Delta f \tag{6.45}$$

$$\overline{i_{\mathrm{s-sp}}} = 2(e\eta G)^2 FP\Delta f / hv \tag{6.46}$$

$$\overline{i_{\mathrm{sp-sh}}} = 2e^2 \eta GF \Delta\nu_{\mathrm{opt}} \Delta f \tag{6.47}$$

$$\overline{i_{\mathrm{sh}}} = 2e^2 \eta GP \Delta f / hv \tag{6.48}$$

where G is the pre-amplifier small signal gain, F is the pre-amplifier noise figure, $\Delta\nu_{\mathrm{opt}}$ is the optical bandwidth occupied by the spontaneous emission and the other terms are as previously defined above. It is easy to show that $\overline{i^2_{\mathrm{s-sp}}} < \overline{i^2_{\mathrm{sh}}}$ and $\overline{i^2_{\mathrm{sp-sp}}} < \overline{i^2_{\mathrm{sp-sh}}}$ and if the filter bandwidth is reduced such that the optical spectrum bandwidth, $\Delta\nu_{\mathrm{opt}}$, is less than 1 nm (125 GHz) then $\overline{i^2_{\mathrm{s-sp}}} < \overline{i^2_{\mathrm{sp-sp}}}$ and the signal-spontaneous beat noise dominates the detection process for high P. The effect of the pre-amplifier on the interferometric phase resolution can be visualized by plotting the carrier powers and power spectral density (assuming the noise sources are white) of the dominant noise sources in the two cases (i.e. with and without a pre-amplifier) as a function of the total received power, P, expressed in units of dB re mW or dBm. This is shown in Figure 6.14 for a pre-amplifier with a small signal gain and noise figure of 100 and 2.5, respectively, and $R_{\mathrm{L}} = 1.28 \times 10^5\ \Omega$, $V = 0.5$, $\eta = 0.8$, $T_{\mathrm{e}} = 600$ K, $\Delta\nu_{\mathrm{opt}} = 125$ GHz, $\Delta f = 1$ Hz and $\lambda_0 = 1550$ nm.

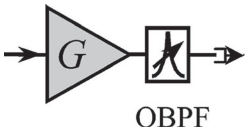


FIGURE 6.13 Optically pre-amplified detector—optical bandpass filter.

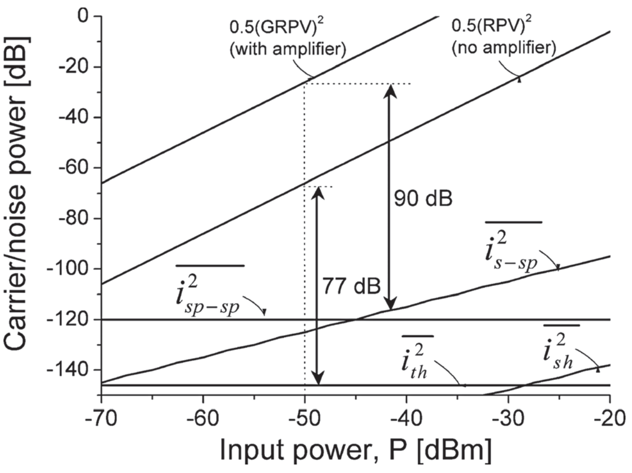


FIGURE 6.14 Signal and noise powers generated with and without a pre-amplifier.

The top traces show the carrier power with the pre-amplifier $(GRPV)^2$, and without, $(RPV)^2$, the difference being due to the pre-amplifier gain, G. The bottom trace shows $\overline{i^2_{\mathrm{sh}}}$ corresponding to the no amplifier case and dominates the CNR when P is greater than –30 dBm. When P is less than –30 dBm, $\overline{i^2_{\mathrm{th}}}$ sets the maximum CNR. With the pre-amplifier, the CNR is set by $\overline{i^2_{\mathrm{sp-sp}}}$ for P greater than −45 dBm. The improvement in phase resolution due to the pre-amplifier is illustrated for a received optical power of –50 dBm (10 nW). Without a pre-amplifier, a maximum CNR of 77 dB re Hz$^{-1/2}$ (200 μrad Hz$^{-1/2}$ using equation (6.41)) is attainable, but with a pre-amplifier, the CNR is increased to 90 dB re Hz$^{-1/2}$ (45 mrad Hz$^{-1/2}$), providing an improvement of 13 dB in the interferometric phase resolution.

6.5.3 Dynamic Range

The dynamic range specifies the difference in maximum and minimum signal amplitude that can be measured by the hydrophone. By assuming an ideal demodulator (i.e. one that will not impose any limitation on the dynamic range or add any excess noise to the signal), the dynamic range can be deduced by consideration of the maximum bandwidth available for the phase-modulated signal to occupy. Carson's rule [67] states that for sinusoidal angular modulation, the bandwidth occupied by the modulation signal is equal to twice the sum of the peak frequency deviation, f_{pk}, and the signal frequency, f_{s},

$$\Delta B = 2(f_{\mathrm{pk}} + f_{\mathrm{s}}). \tag{6.49}$$

Since

$$\phi(t) = \phi_0 \sin(2\pi f_{\mathrm{s}} t) \text{ and } \omega_{\mathrm{pk}} = \left| \frac{\mathrm{d}\phi(t)}{\mathrm{d}t} \right|$$

then

$$\left| f_{\mathrm{pk}} \right| = \phi_0 f_{\mathrm{s}}$$

and

$$\Delta B = 2 f_s (\phi_0 + 1). \tag{6.50}$$

Thus, for a given signal frequency, the maximum modulation depth (and, hence, signal amplitude) can be determined from knowledge of the available bandwidth. This is then combined with the interferometric phase resolution to determine the dynamic range. It therefore remains to determine the available bandwidth for each interrogation scheme. In the heterodyne scheme, the available bandwidth is simply $2\Delta\omega/2\pi$ (i.e. twice the carrier frequency). In the PGC scheme, the available bandwidth is given by $\omega_{pgc}/2\pi$, as illustrated in the inset of Figure 6.11. By comparing equations (6.31) and (6.37), it is apparent that the PGC signal occupies more bandwidth than the heterodyne signal. Hence, the PGC scheme requires more bandwidth than the heterodyne scheme to achieve a given dynamic range. We will show in the next section the dynamic range one would expect from a multiplexed sensor.

6.5.4 Digital Techniques

Early hydrophone systems implemented phase demodulators based on the analogue differentiate and cross-multiply techniques for the homodyne PGC scheme and frequency discriminators or phase-locked loops for the heterodyne scheme. The ubiquity of low-cost, high-speed digital samplers has enabled dramatic improvements in performance and miniaturization to be obtained. Direct sampling of the photodiode signals and implementation of the signal processing methods described in the previous section have enabled the dynamic range to become limited by the sample rate alone by implementing fringe counting algorithms to track phase excursions exceeding 2π. Phase resolutions limited by optical noise sources can be obtained with sampling resolutions greater than 12-bit. This processing can be implemented into field-programmable gate arrays (FPGA) enabling real-time demodulation of multiplexed sensor systems. The ability to demodulate signals from hundreds of sensors in a single FPGA has been demonstrated.

6.6 Optical Systems and Multiplexing

A multiplexing scheme must be designed with the following objectives: (i) to maximize the number of sensors addressable through a small number of fibres; (ii) to maximize the efficiency with which power is delivered (returned) to (from) each sensor; (iii) to minimize the sensor to sensor crosstalk; (iv) to maintain comparable performance from a multiplexed sensor to that of a single sensor and (v) to maintain an electrically passive array. Reviews of multiplexing schemes can be found in [68,69], which cover schemes based on frequency, time, wavelength and coherence-based approaches. Here, we present two techniques that have been demonstrated to achieve these requirements. The first technique to be discussed is known as frequency domain multiplexing (FDM) with homodyne PGC. We shall then focus on time division multiplexed (TDM) architectures followed by a discussion of recent work on large-scale multiplexing architectures. Techniques to overcome polarization fading are then considered, followed by a discussion on new technology for fibre optic acoustic sensors based on fibre laser sensors.

6.6.1 The FDM Technique

This technique is based on frequency multiplexing of carrier frequencies and has demonstrated the potential to multiplex up to 64 sensors onto 16 fibres using eight semiconductor LD sources. This approach was implemented with the homodyne PGC interrogation scheme in a matrix-type architecture [8]. The array is driven by a series of LDs, each modulated at a different frequency. The matrix is composed of $N \times N$ sensors and N sources, as shown in Figure 6.15a for the case of a 2×2 matrix. Each source interrogates N sensors and the outputs from the sensors are combined so that the received signal produces N carrier frequencies on the photodiode, each corresponding to a single sensor. This scheme has been shown to achieve laser-phase-noise-limited detection sensitivity and sensor to sensor crosstalk of less than −60 dB and has the advantage of interrogating all the sensors continuously.

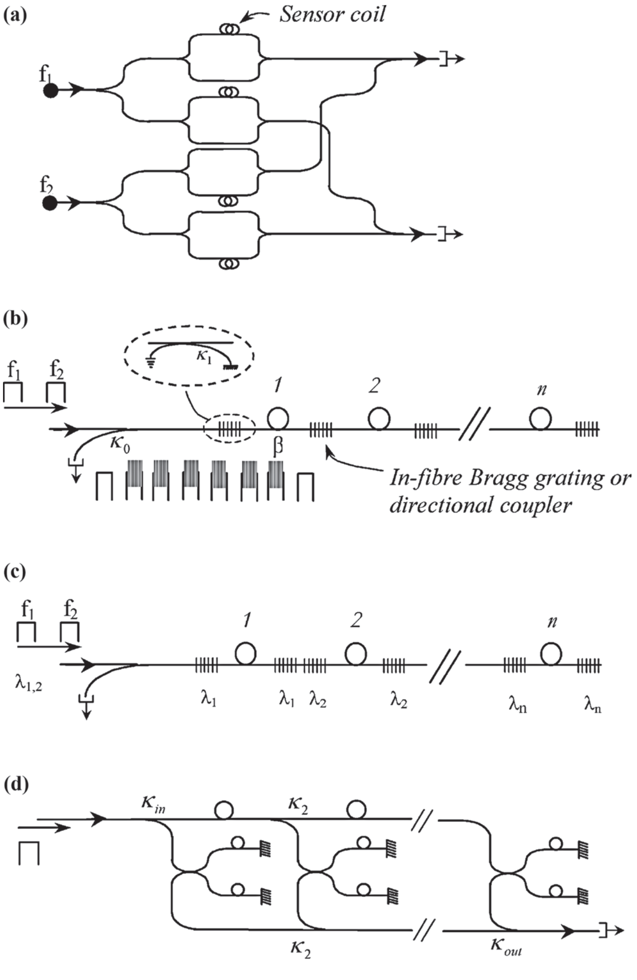


FIGURE 6.15 Multiplexing architectures: (a) PGC FDM 2×2 matrix [8], (b) TDM reflectometric/in-line Michelson [73], (c) TDM/WDM reflectometric [76] and (d) TDM progressive ladder, Michelson.

However, the large number of telemetry fibres required renders this technique less appealing than the TDM methods described next, for many applications.

6.6.2 TDM Architectures

TDM has been demonstrated to multiplex many sensors onto a single fibre and achieve most of the requirements described earlier. Schemes using both heterodyne and PGC interrogation have been demonstrated to multiplex up to 64 hydrophones on two fibres [71,72]. Two architectures based on this scheme are now discussed.

The pulsed reflectometric architecture. This architecture, shown in Figure 6.15b, consists of serially multiplexed Fabry–Pérot interferometers, with low reflectivity mirrors that can be interrogated by two frequency-shifted pulses separated in time by an amount equal to twice the transit time in the interferometer arm, 2τ. The received signal consists of a series of pulses generated from reflections of the first interrogating pulse overlapped in time but delayed by 2τ with the pulse train generated from the reflections of the second interrogating pulse. Where overlapping of pulses occurs, an interference signal or heterodyne signal is generated. Each sensor generates its own heterodyne signal upon which the modulating signal of interest will appear [73]. Early development of this topology used semi-reflecting splices to produce the reflections; however, multipath interference induced severe cross-talk that limited the number of sensors that can be multiplexed onto a fibre [74]. To overcome this problem, an asymmetric reflector was developed using a directional coupler that has a reflective mirror formed on one of the output ports. By index-matching the unused input port, the device will produce a reflection only in one direction, as shown in the inset of Figure 6.15b. In this configuration, the architecture is known as the in-line Michelson. To further improve the optical power budget of this type of array, the coupling coefficients of the couplers are set such that equal power is returned from each sensor at the photodiode.

If the individual losses (coupler excess loss, two splices, fibre attenuation) are lumped into a single element, β, the mirror loss is α, the total number of sensors is N and κ_0 is the coupling coefficient of the telemetry coupler (usually 50%), then the power returned from the first mirror for a single injected pulse of power, P_{inc}, is

$$P_1 = \kappa_0^2 \kappa_1^2 \alpha^2 P_{\text{inc}} \tag{6.51}$$

where κ_j is the coupling coefficient of mirror j. Thus, the power returned from the jth mirror is

$$P_j = \kappa_0^2 \alpha^2 \prod_{i=1}^{j-1} (1-\kappa_i)^2 \beta^2 P_{\text{inc}}. \tag{6.52}$$

From the condition that $P_j = P_{j+1}$,

$$\kappa_j = \frac{\beta \kappa_{j+1}}{(1+\beta \kappa_{j+1})} \tag{6.53}$$

and a closed-form expression for κ_j can be determined using the condition, $\kappa_{N+1} = 1$,

$$\kappa_j = \frac{1-\beta}{\beta^{j-N-1} - \beta}. \tag{6.54}$$

For a loss-less system ($\beta = 1$), (6.54) simplifies to

$$\kappa_j = \frac{1}{N-j+2}. \tag{6.55}$$

The insertion loss of the network is defined as the ratio of the peak pulse power returned from any sensor, P_n, to the injected peak pulse power, P_{inc}, and, assuming $\kappa_0 = 0.5$, is

$$IL_{\text{PRA}} = \frac{P_n}{P_{\text{inc}}} = \frac{1}{4(N+1)^2}. \tag{6.56}$$

This architecture has been shown to achieve average sensor–sensor crosstalk levels of less than −47 dB [75], limited by the finite extinction ratio of the drive signal for the AOM generating the optical pulses.

A variant of this architecture using in-fibre Bragg gratings (IFBGs) as reflective elements and including WDM to overcome the cross-talk problem described earlier has also been demonstrated. An array consisting of four serially multiplexed sensor coils, each with a pair of IFBG reflectors, where every other sensor operates at the same wavelength was demonstrated, as shown in Figure 6.15c. By incorporating a wavelength filter at the array output, cross-talk between sensors operating at the same wavelength was shown to be less than –30 dB and between sensors operating at different wavelengths was less than −60 dB [76].

The progressive ladder architecture (PLA). This architecture, shown in Figure 6.15d incorporating Michelson interferometers, comprises separate fibres for the launch and return pulses. Each sensor forms an individual Michelson interferometer with a small path imbalance that allows interrogation with the homodyne PGC technique. The inclusion of reference coils for each interferometer confines the sensitive region of the sensor to within the interferometer (i.e. the leads connecting the sensors are not part of the interferometer arms). Another advantage of the PLA is that the couplers that occupy a symmetric location with the network have the same coupling coefficient. Closed-form expressions for the coupling coefficients for a loss-less network have been derived in [31] and we quote the results here. The coupling coefficient for the input and output couplers is given by

$$\kappa_{\text{in}} = \kappa_{\text{out}} = \frac{\kappa^2}{\kappa^2 - \kappa + 1} \tag{6.57}$$

and the optimum value of coupling coefficient for the couplers positioned symmetrically within the network (i = 2, 3, . . . , $N-1$) is given by

$$\kappa = \kappa_i = \frac{1}{3} - \frac{0.84(N+3)}{A(N-3)^{1/2}} + \frac{0.26A}{(N-3)^{1/2}} \tag{6.58}$$

where

$$A = [39(N-3)^{1/2} - 7N(N-3)^{1/2} + 3^{3/2}(3N^3 - 15N^2 + 149N - 137)^{1/2}]^{1/3}. \tag{6.59}$$

For example, for a loss-less network, optimum coupling coefficients are $\kappa_{\text{in}} = \kappa_{\text{out}} = 0.05$ and $\kappa_i = 0.2$ for $N = 10$. Closed-form expressions for the lossy network are not trivial to derive and a numerical analysis is presented in [31]. This shows that, assuming typical losses, the required coupling coefficients deviate only slightly from that predicted by (6.58). It can also be shown that the insertion loss, assuming a loss-less network, is well approximated by

$$IL_{\text{PLA}} \approx \frac{1}{N^2}. \qquad (6.60)$$

Thus, the power returned from a sensor in the PLA is a factor of approximately four higher than in the PRA. Characterization of sensor–sensor crosstalk in this network has not been published; however, a similar architecture has been shown to achieve a sensor-sensor crosstalk of less than −50 dB [77,78].

6.6.3 TDM Architectures Analysis

For both of the architectures described earlier, the number of sensors that can be multiplexed onto a fibre is limited by the required interrogation rate for each sensor (determined by the dynamic range requirements discussed later) and the optical power. For example, the maximum interrogation rate, f_{rep}, can be deduced from the transit time of the optical pulse through the array, which in the case of the PRA is $1/2\tau(N+1)$ Hz. Assuming the pulse width is set to 2τ, the optical duty cycle, D, is $1/(N+1)$.

It is now possible to calculate a power budget for a TDM array, and the launch optics for a typical system is shown in Figure 6.16. This can be used to interrogate an in-line Michelson array of the type shown in Figure 6.15b. The path-balancing unit generates two optical pulses, separated in time by 2τ (determined by the delay coil length) with a frequency difference of $\Delta\omega$, from the same time segment of light. This ensures that the effective path imbalance in each interferometric sensor is minimized, thus minimizing laser-frequency-induced phase noise. For an array consisting of eight sensors ($N = 8$) and a 5 km datalink length, the power budget is given in Table 6.4, which shows the power returned from a sensor assuming only a single pulse is launched. The shot-noise-limited phase resolution can then be calculated from equations (6.41) and (6.42), where P is substituted with the average power returned from a sensor, given by $2\,DP_{\text{n}}$. The factor of two accounts for the fact that we now have an optical power, P_{n}, returned from each interferometer arm, rather than $P_{\text{n}}/2$.

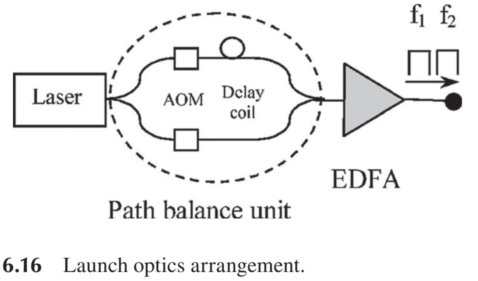


FIGURE 6.16 Launch optics arrangement.

TABLE 6.4

Example Power Budget

	Power Budget	Unit
Laser power	10	dBm
Path balance unit IL	–9	dB
Power EDFA gain	20	dB
Datalink IL, 5 km (two-way) (0.3 dB km^{-1})	–3	dB
PRA array IL, ($N = 8$)	–25	dB
Margin	–5	dB
Received single-pulse power, P_{n}	–12	dBm

Taking $r = 1$A W^{-1}, $V = 0.5$ and $D = 1/9$, the phase resolution, $\delta\phi$, is about 3 µrad Hz$^{-1/2}$. In practice, the other noise sources described in Section 6.5 will reduce the phase resolution but a resolution between 10 and 30 µrad Hz$^{-1/2}$ is realistically achievable with careful choice of component specification. Thus, for a hydrophone responsivity of 0.5 rad Pa^{-1} and a phase resolution of 10 µrad Hz$^{-1/2}$, the pressure resolution would be about 20 µPa Hz$^{-1/2}$.

Assuming the interferometric phase resolution is 10 µrad Hz$^{-1/2}$, then using equation (6.28), the strain resolution in 100 m of fibre is, therefore, about 2×10^{-14}.

The dynamic range of a sensor in this multiplexing architecture can be deduced by considering the frequency spectrum of a single heterodyne pulse and determining the available bandwidth. We consider the case for $f_{\text{rep}} \ll f_{\text{c}}$. The frequency spectrum in this case is shown in Figure 6.17. To maximize the available bandwidth, the condition $f_{\text{c}} = f_{\text{rep}}/4$ should be met so that the spacing of adjacent carrier components is maximized. In this case, $\Delta B = 2f_{\text{c}}$ where $f_{\text{c}} = \Delta\omega/2\pi$. Using the example given earlier, taking the length of fibre in a hydrophone $L_{\text{s}} = 100$ m and using the relation $\tau = nL_{\text{s}}/c$ gives $f_{\text{rep}} = 111$ kHz. Thus, $f_{\text{c}} = 28$ kHz, $\Delta B = 56$ kHz and using equation (6.50) gives $\phi_0 = 27$ rad. Taking the phase resolution as 10 µrad Hz$^{-1/2}$, the dynamic range is 129 dB re Hz$^{-1/2}$. In a hydrophone array, this dynamic range would be reduced slightly from the increase in noise from ambient acoustic noise in the ocean. More detailed analysis of these systems can be found in the literature [75].

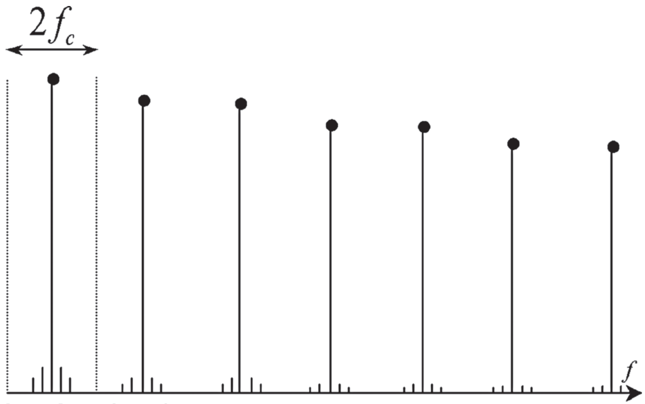


FIGURE 6.17 Spectral content of heterodyne pulse for case when $f_{\text{rep}} <\ll f_{\text{c}}$ (DC term removed).

6.6.4 Large-Scale Array Architectures

Full-size arrays may contain over 1000 sensors; therefore, efficient methods of combining more sensors onto one fibre have been investigated. A technique based on incorporating Er^{3+}-doped fibre amplifiers in the rungs of a ladder architecture to overcome the power splitting losses has been investigated by two groups [79–82] and an example of such an architecture is shown in Figure 6.18a. An experimental system has been set up with a ten-rung ladder and 20 Er:Yb-doped fibre amplifiers. By considering the effective noise figure of such an array compared to that of a standard ladder architecture, a system capable of driving up to 300 sensors whilst maintaining an interferometric phase resolution close to 1 μrad $Hz^{-1/2}$ was demonstrated [82].

Another technique based on combining DWDM and TDM has been demonstrated with the aim of increasing the number of sensors addressable through a fibre pair. The architectures use optical add/drop multiplexers (OADM) to couple a wavelength from a telemetry fibre into a TDM array module (made up of a ladder architecture or PRA) and then either re-combine the return signal back onto the telemetry fibre or onto a separate fibre as shown in Figure 6.18b. An experimental system was demonstrated using a three-wavelength Er^{3+}-doped fibre laser source and heterodyne detection with in-line Michelson-type array modules. The system demonstrated crosstalk between a sensor in one array module and the corresponding sensor in an array module on an adjacent wavelength to be less than −72 dB and was limited by the finite adjacent channel isolation of the wavelength-demultiplexer on the array output. The system was demonstrated using two wavelengths to interrogate 12 sensors; however, by using 16 wavelengths at a spacing of 1.6 nm within the Er^{3+} window, more than 500 sensors could be interrogated through two fibres [75,83].

6.6.5 Overcoming Polarization-Induced Signal Fading

When two coherent beams from an interferometer interfere, the fringe visibility, *V*, is dependent on the relative states of polarization (SOPs) of the two beams. Maximum visibility occurs when the SOPs are aligned and no interference occurs when the SOPs are orthogonal. It is well known that standard telecom grade optical fibre exhibits a small birefringence that causes the SOP of the propagating beams to change randomly as it travels along the waveguide. Also, external effects such as asymmetric stresses (due to changes in pressure or temperature) or twist of the fibre will affect the birefringence of the fibre [28]. These environmental factors result in an effect known as polarization-induced signal fading in interferometric sensors, where the CNR varies with time due to a time-varying fringe visibility. In some circumstances, a finite degradation in the CNR can be tolerated; however, some form of polarization control scheme is needed to avoid complete extinction of the interference signal. An elegant method to achieve maximum fringe visibility from Michelson interferometers uses a type of ortho-conjugate mirror known as an orthogonal mirror [63]. This device makes use of the Faraday effect to produce a reflected beam that has a SOP orthogonal to that of the incident beam. If used in place of the standard mirrors in a Michelson interferometer, the light received from each arm will always be in a polarization state orthogonal to that of the

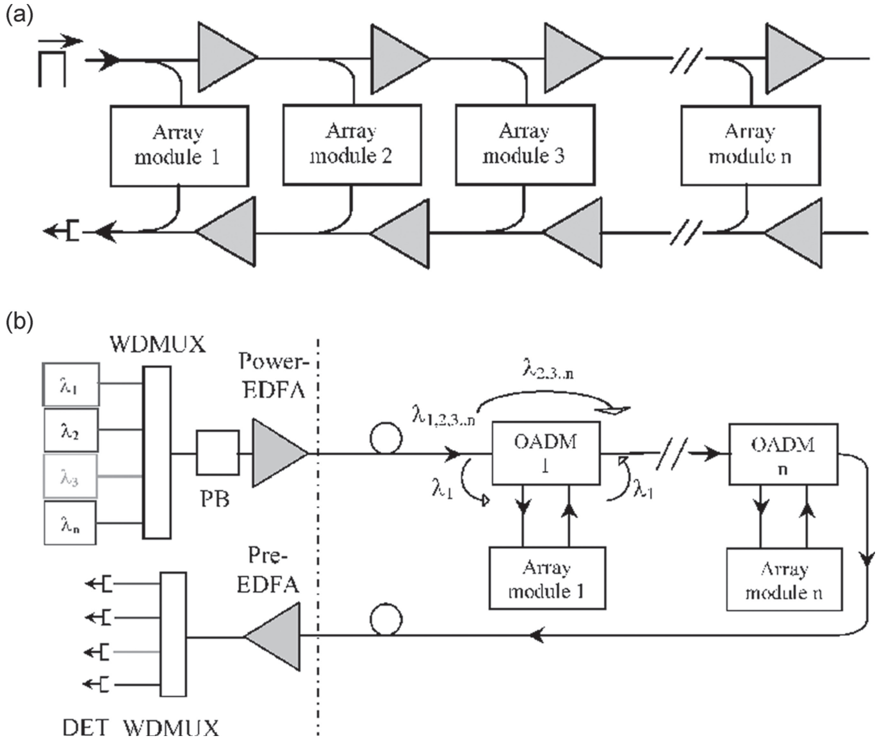


FIGURE 6.18 Large-scale multiplexing architectures: (a) ladder architecture with EDFA telemetry after [82] and (b) DWDM and TDM architecture after [75].

launched light, regardless of the birefringence properties of the fibre interferometer. The non-reciprocity of the fibre birefringence for counter-propagating beams effectively untwists the effect on the SOP of the birefringence on the journey to the mirror during its return journey. These devices have been used in interferometric sensors and have shown that maximum visibility can be maintained from multiplexed sensors [84]. However, although these devices are electrically passive, their high cost usually precludes their use in large arrays of sensors. The use of high birefringence fibre has also been attempted but component costs and array construction complexity tends to render this approach unviable.

The polarization and birefringence properties of fibre interferometers have been studied extensively, and several schemes have been proposed that rely on either control of the input light SOP or SOP selection of the received light. These are now described.

Techniques based on input-light polarization control. Analysis of the polarization properties of fibre interferometers revealed that the fringe visibility was dependent on both the input SOP of the light to the interferometer and the birefringence properties of the two fibre arms. It was shown that there exist unique SOPs or eigenmodes of the polarization for which optimum visibility was obtained and this formed the basis of a technique that used active feedback to maintain optimum visibility [85]. Another technique that used pseudo-depolarization of the input light to the interferometer to overcome fading due to the birefringence effects of the input lead has also been demonstrated [86]. These techniques, however, cannot fully overcome polarization-induced signal fading for multiplexed arrays. However, more recently, a technique based on switching of the polarization state of the input light into an array of sensors has been demonstrated [87]. The principle is shown in Figure 6.19 for an in-line Michelson array. An electro-optic polarization modulator is placed between the laser source and AOM that allows the linear SOP of the laser output to be rotated by 90° and the SOP of the second pulse in alternate pulse pairs to be rotated by 90° relative to the others. When the fringe visibility from one set of pulses falls to zero, the fringe visibility of the second set of interrogating pulses will be approaching maximum. A selection circuit is then employed to choose the interference signal with the largest visibility. Successful operation of this technique relies on the fact that orthogonally polarized light launched into a fibre will remain orthogonal in polarization regardless of the birefringence properties of the fibre and assumes that there is no polarization-dependent loss [88].

Techniques based on received-light SOP selection. If the SOP of the two beams from the arms of an interferometer becomes orthogonal, then direct detection will produce no interference. However, by passing the light through a linear polarizer, prior to detection, that is orientated such that there exists a projection of both polarization states onto the axis of the polarizer; the transmitted light will produce an interference signal since both SOPs are now aligned with the polarizer axis. This breaking of the orthogonality between the SOP of the light beams forms the basis of a technique known as polarization diversity detection and has been demonstrated using a mask formed from three linear polarizers with their axes aligned 60° to each other. The light from the interferometer output fibre is expanded and passed through the mask and the output of each polarizer is detected separately. It was shown that, for all input SOPs, there exists an output that exhibits a non-zero fringe visibility. For slow variations in fringe visibility, selection of the signal with the highest fringe visibility allows for continuous operation of the sensor [89,90].

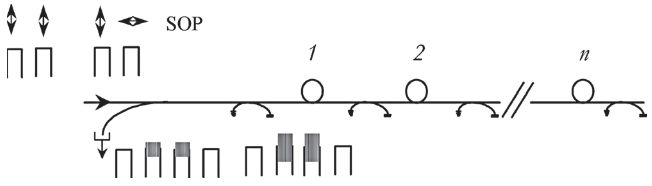


FIGURE 6.19 Polarization switching technique with the in-line Michelson architecture after [87].

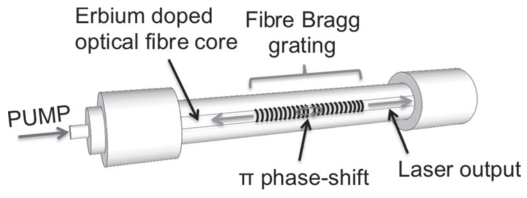


FIGURE 6.20 FBG laser sensor.

6.6.6 Future Trends in Optical Hydrophone Technology—Fibre Laser Sensors

Optical hydrophones have reached a level of development whereby arrays have been deployed in real-world environments and shown to achieve the required performance. This drives the development of new optical-fibre-based technologies that may provide advantages over the existing technologies described earlier. One such technology is that based on the IFBG and the in-fibre laser (IFL). This device consists of a Fabry-Perot cavity formed with two closely spaced FBGs at identical wavelength or a long FBG with a pi phase shift in the centre, as illustrated in Figure 6.20. By forming the Bragg gratings in an active fibre (such as erbium-doped fibre) that provides optical gain, lasing can be achieved. An array of lasers can be made by forming Bragg grating pairs at different wavelengths along a single fibre, which can be pumped with a single semiconductor LD. Changes in the cavity length of each laser are encoded on the respective laser frequency, which can be decoded using interferometric techniques [91]. The strain resolution of such a laser is found to be limited by thermodynamic noise within the laser cavity yielding fundamentally limited strain resolutions less than 10^{-13} $\varepsilon/\text{Hz}^{1/2}$ [92]. These sensors have been utilized in several hydrophone applications [93–95].

6.7 The Optical Geophone

In this section, we describe the optical geophone, which is a fibre optic sensor developed to measure vibration rather than pressure changes, and in particular to measure ground motion.

Such sensors have found application in the oil and gas industry for seismic surveying, as is discussed further in Section 6.8.

Development of the optical geophone accelerated in the late 1990s, from a number of sources including industrial and military accelerometers. A considerable design effort led to the development of sensors with the required performance characteristics for seismic use, and in the early 2000s the first field tests of such systems took place. Parallel development in multiplexing and associated techniques led to the eventual realization of commercially viable systems which remain amongst the largest optical sensing systems ever constructed, as is described below.

6.7.1 The Basic Transduction Mechanism

A geophone is a sensor which is designed to be sensitive to motion (acceleration, velocity or displacement) along one axis (note that sometimes in the seismic industry, the term geophone is used for a sensor which operates above resonance, and has a response to velocity that is independent of frequency, while the term accelerometer is used for a device which operates below resonance, and has a response to acceleration that is independent of frequency. Here, however, we use geophone in its wider sense as a sensor to measure ground motion.

The majority of optical geophones are interferometric sensors, in which the sensing element is a length of optical fibre, typically arranged in a coil in the length range of 10–50 m. In this respect, they have much in common with an optical hydrophone, but in this case the coil is mechanically packaged such that motion along the axis of the coil (most usually acceleration) is converted into a liner phase change in an optical signal passing through the coil. This is achieved by converting the motion into a strain in the fibre, which causes a phase change through a combination of physical length change and the elasto-optic effect. Typically, the coil is also packaged in such a way to minimize effects of transverse motion (in this respect, the geophone differs from a hydrophone: a hydrophone is designed as an omni-directional sensor, while the geophone is normally designed to be only sensitive to motion in one direction).

In practice, most optical geophones are actually accelerometers and are designed to be used below resonance. The resonance frequency of such devices varies but is typically in the range 500 Hz to 2 kHz. As with any mass-spring device, the resonance is proportional to the stiffness of the spring and the sensitivity is inversely proportional to the stiffness so there is normally a trade-off between sensitivity and useable bandwidth.

6.7.2 Alternative Geophone Designs

Geophone design is typically aimed at maximizing the sensitivity and bandwidth of the sensor along its intended operational axis, while minimizing its sensitivity in other axes, and to achieve this with high linearity and wide signal range. Other likely design considerations relate to the operating environment. There are 3 main environments: subsea (usually on the seabed), surface land and borehole (underground oil wells). The borehole environment is especially demanding as it requires operation at temperatures of over 150°C, as well as isolation from high pressure and large amounts of hydrocarbon. Geophones have been produced based on rubber or metal mandrel designs, and on flexural disks. We describe each of these designs briefly below.

Some of the earliest geophone designs were based on a rubber mandrel, which offer a convenient transduction mechanism for converting an axial motion into a phase change in an optical fibre coil [103]. This is achieved because of the high incompressibility of rubber, such that an axial motion if the runner mandrel is converted into a radial expansion, which directly strains an optical fibre wound onto the mandrel (Figure 6.21).

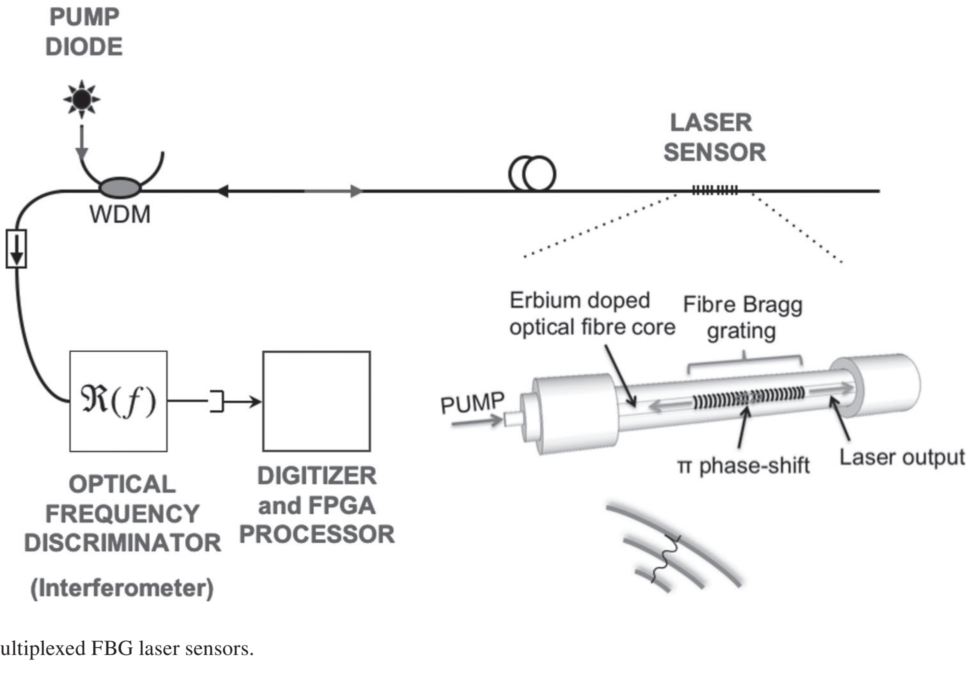


FIGURE 6.21 Multiplexed FBG laser sensors.

A later version of the rubber mandrel design, fully commercialised for geophysical use, is described in [96]. The sensor is shown in Figure 6.22 and a schematic of its operation is given in Figure 6.23. The sensor uses 40 m of optical fibre coiled around a rubber mandrel. It has a resonance frequency of 830 Hz and a phase responsivity of 60 dB re 1 rad/g. The cross-axis responsivity ratio is greater than -50 dB.

Another class of mandrel-based geophone uses a mandrel comprising a thin metal shell, which may be cylindrical or elliptical in form. Such sensors were used in the first large-scale deployment of a seabed geophysical array, as described in 6.8. An alternative geophone design makes us of a fibre coil attached to a flexural disk [104]. Such designs were originally used in the US for military applications, but more recently have been sued on other configurations for seabed geophysical applications.

6.7.3 System Configurations for Geophone Use

Geophysical applications require large, complex arrays which may have well been in excess of 10 000 sensor channels. Other requirements include an operational bandwidth of less than 0.5 Hz to greater than 500 Hz, a low noise floor (typically less than 100 ng/Hz$^{1/2}$) and a dynamic range of at least 120 dB. This is a demanding set of requirements, which has required advances in system architecture and enabling technologies, in parallel with the geophone development described above.

These advances have included the introduction of very high density WDM techniques, which have allowed up to 512 sensor channels per fibre pair [97]. Wavelength discrimination can take place either at the sensor, using FBGs, or at separate multiplexing points within the architecture. Further extension of multiplexing to as much as 3000 channels per fibre pair can also be achieved with the use of remote and distributed optical amplification. Other enabling technologies have included very narrow linewidth lasers optimized for low-frequency performance and the use of Faraday mirrors to remove polarization effects. Highly efficient digital demodulation algorithms and processing schemes have been implemented to greatly increase the efficiency of the data-handling process.

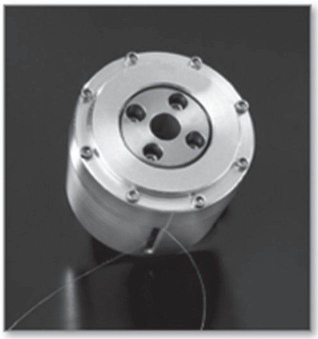


FIGURE 6.22 Optical geophone.

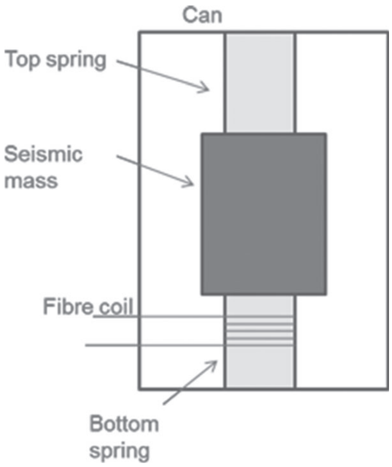


FIGURE 6.23 Optical geophone schematic.

6.8 Application Studies

In this section, we describe field demonstrations of prototype multiplexed fibre optic sensor system. The first at-sea deployment of an optical hydrophone array was reported by Plessey, UK, in 1986 [11] which described a six-element array deployed in shallow water off the south coast of the UK. The array consisted of six mandrel-based hydrophones attached to a 12-m long support gantry and lowered into approximately 20 m of water. The hydrophones were arranged in a pulsed reflectometric architecture and interrogated with a He-Ne laser operating at 1.15 μm. Despite the limitations of the component technology at the time, the system demonstrated a pressure resolution of 44 dB re μPa Hz$^{-1/2}$ at 500 Hz limited by excess electronic demodulator noise and a dynamic range of 90 dB.

A more recent demonstration comprised 96 hydrophones combined using time and WDM. This system was field-tested during a sea trial in 2002 and used to track synthetic acoustic targets [98,99]. The hydrophone used in this system is based on a fibre-wrapped air-backed mandrel design [35]. The physical change in the diameter of a plastic mandrel, around which the fibre is wrapped, under the influence of a time-varying pressure field induces a strain in the fibre. An omni-directional hydrophone response is obtained when the acoustic wavelength is much greater than the maximum hydrophone dimension. Multiplexing is achieved by serial concatenation of sensors and using the time of flight of injected optical pulses to sequentially address each sensor. WDM is incorporated with TDM to permit signals from several TDM sensor arrays to be combined onto a single optical fibre. In the prototype system, 16 sensors are multiplexed with time (expandable to 64) and six wavelengths are used to allow 96 hydrophones to be interrogated through two optical fibres. To achieve high-phase resolution from the interferometric sensor, a high-coherence laser is required that emits a stable single optical frequency. Six erbium-doped distributed feedback fibre lasers are used in this system with emission wavelengths ranging from 1541.35 nm to 1549.32 nm spaced by 1.6 nm. A schematic of the system layout is shown in Figure 6.24. It can be separated into three subsystems: the launch optics, hydrophone array and the receive optics and electronics. The launch and receive optics are usually co-located. The launch optics generates and amplifies two

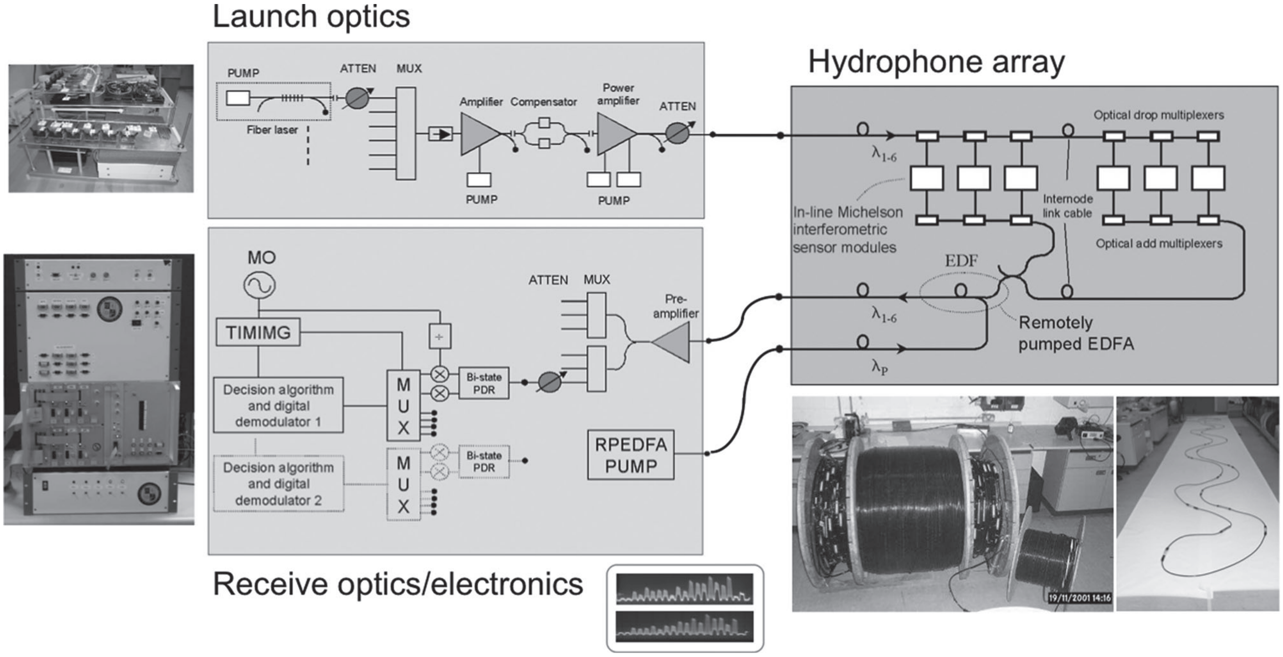

Launch optics
PUMP
ATTEN
MUX
Fiber laser
Amplifier
Compensator
Power amplifier
ATTEN
PUMP
PUMP
Hydrophone array
Optical drop multiplexers
λ_{1-6}
In-line Michelson interferometric sensor modules
Internode link cable
Optical add multiplexers
EDF
λ_{1-6}
Remotely pumped EDFA
λ_P
MO
TIMIMG
ATTEN
MUX
Pre-amplifier
Decision algorithm and digital demodulator 1
M U X
Bi-state PDR
Decision algorithm and digital demodulator 2
M U X
Bi-state PDR
RPEDFA PUMP
Receive optics/electronics
19/11/2001 14:16

FIGURE 6.24 System arrangement of multiplexed hydrophone array.

pulses at each wavelength. The maximum launch power at each wavelength is limited by stimulated Brillouin scattering, which occurs at about 10 mW of average power in conventional single-mode optical fibre. The hydrophone array contains the sensors and wavelength multiplexers. Each wavelength is coupled-off sequentially using optical add-drop multiplexers and injected into a reflective in-line Michelson array, of the type illustrated in Figure 6.19 using directional couplers. The return pulses are combined onto a return fibre. To allow the stand-off distance to be increased, a remotely pumped EDFA is incorporated into the array, providing around 20 dB of gain to the return signals. The 1480 nm pump, λ_p, for this EDFA is provided through a separate fibre. The receive electronics contains an EDFA to amplify the return signals before separating them through a wavelength de-multiplexer followed by detection. The detector consists of a polarization diversity receiver, which overcomes the problem of polarization-induced signal fading [89], and the remaining electronics performs the trigonometric calculations to recover the phase from each sensor.

The 96 hydrophones are separated into two arrays of 48 hydrophones. These were deployed a few kilometres from the coast of a major military and commercial shipping port. A 3 km fibre-optic cable joined each array and a 5-km cable connected array 1 to the shore station. For the majority of the tests, an extra 35 km of optical fibre wound onto spools was added to the input and output fibres, to increase the effective array stand-off to 40 km. Array 1 was deployed at a depth of 57 m and array 2 at a depth of 73 m. The ambient acoustic-induced phase noise was found to be approximately 20 dB higher than the mean sensor self-noise at ~488 Hz for array 1 and up to 10 dB higher than the sensor self-noise for array 2. The temporal outputs of hydrophones from array 1 without high-pass filtering are shown in Figure 6.25. The vertical striation represents the slow-pressure variations that propagate along the hydrophone array due to surface wave motion. This plot illustrates the capability of fibre optic hydrophones to resolve very low-frequency variation in ambient pressure.

Characterization of the beam-forming performance of the array was achieved with a towed acoustic source. The source transmitted a number of discrete frequencies, and since the angular resolution of an array increases with frequency, tracking is most accurately achieved using the highest of the discrete frequencies, which is 445 Hz. To show the capability of the system to track the towed source only the section of the beam pattern at 445 Hz is used, and this is combined with the same section from subsequent beam patterns. This enables an image to be created that shows how the bearing and strength of the signal at 445 Hz vary as the towed source moves. Such an image, shown in Figure 6.25, has been created from 4900 s of data from array 1.

At times up to 2500 s, the strongest signal detected is from the towed source; however, as it moves further away, its signal strength decreases and the signal from another nearby vessel becomes higher.

There are normally two maxima on the trace due to the left/right ambiguity. When the towed source moves near the end-fire position (which happens around 1200 s), the angular resolution of the beam pattern becomes worse and the two traces merge. The right-hand trace is narrower and more intense which shows that this represents the true bearing. The dotted white line on this figure shows the actual bearing of the towed source calculated from the log of its GPS positions. The numbers beside this line show the range in kilometres of the towed source from the array at that time. There is clearly very good agreement between the actual and measured bearing out to a range of around 5.5 km (3000 s) after which time the signal

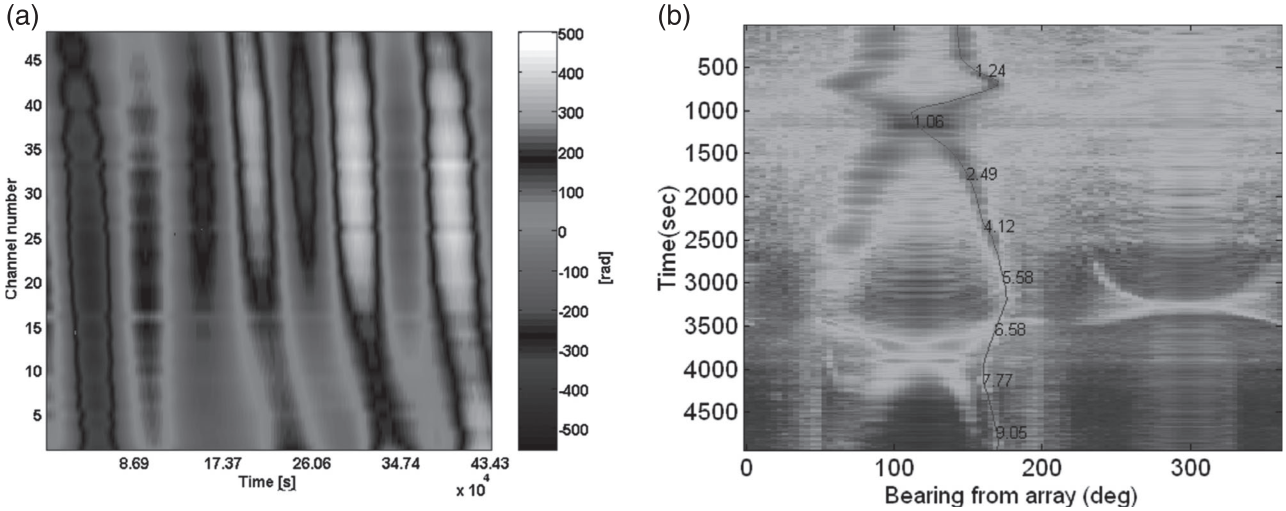


FIGURE 6.25 (a) Temporal output from a 48-sensor array and (b) bearing time plot from a 48-element array.

from another vessel temporarily masks the signal from the towed source. After 4200 s, when the other vessel has moved away, there is still a faint trace that follows the dotted line, which shows that the towed source is being detected to a range of ~9 km.

The first field demonstrations of seabed seismic arrays for geophysical applications took place in the early 2000s. Early demonstrations were of field-prototype systems with typically around 30 channels, demonstrating the sensor performance and the basic system operation. However, much larger systems were soon under construction, and in 2008 a full-scale commercial system was deployed on the Ekofisk oil field in the North Sea. This seabed system was designed to monitor the Ekofisk field during its later production stages. This system, which was based on the mandrel accelerometer, and a TDM/WDM scheme based on FBGs, involved the deployment of 200 km of cable holding 16,000 optical sensing channels (12,000 geophones and 10,000 hydrophones). This system, which was deployed with 100% of channels working, remains the largest field installation of an optical sensor system to date. So far, the system has been in use for 9 years and remains in full operation [105].

Other field tests have been on a smaller scale but have involved some interesting technical developments. In 2011, a 35-km system was deployed at the Jubarte field off Brazil. This used a sensing approach based on a flexural disk geophone, multiplexed using a combination of TDM and WDM. This system was deployed in a water depth of over 1300 m, the deepest water deployment of an optical seismic system yet achieved.

Another approach makes use of a mandrel geophone and hydrophone, in conjunction with a dense TDM/WDM approach [100]. This has been shown to achieve a multiplexing rate of 512 sensors/fibre pair, as well as a dynamic range in excess of 180 dB. This approach has been tested in a field test off Norway in 2008. Figure 6.26 shows this system being deployed in 270 m of water.

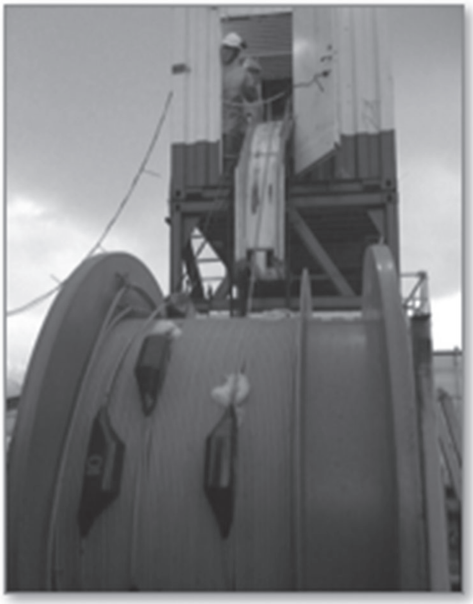


FIGURE 6.26 Optical seismic array being deployed off Norwegian coast.

REFERENCES

1. Stansfield D. 1991 *Underwater Electroacoustic Transducers* (Bath: Bath University Press).
2. Wenz G. M. 1962 Acoustic ambient noise in the ocean: spectra and sources. *J. Acoust. Soc. Am.* **34** 1936–56.
3. Bucaro J. A, Dardy H. D and Carome E. 1977 Fibre optic hydrophone *J. Acoust. Soc. Am.* **62** 1302–4.
4. Cole J. H., Johnson R. L. and Bhuta P. B. 1977 Fibre optic detection of sound *J. Acoust. Soc. Am.* **62** 1136–8 see also Culshaw B., Davies D. E. N. and Kingsley S. A. 1977 Acoustic sensitivity of optical-fibre waveguides *Electron. Lett.* **13** 760–1.

5. Giallorenzi T., Bucaro J., Dandridge A., Sigel G., Cole J., Rashleigh S. and Priest R. 1982 Optical fiber sensor technology *IEEE J. Quantum Electron.* **QE-18** 626–65.
6. Lagakos N., Litovitz T., Macedo P., Mohr R. and Meister R. 1981 Multimode optical fiber displacement sensor *Appl. Opt.* **20** 167–8.
7. Rashleigh S. C. 1980 Acoustic sensing with a single coiled monomode fiber *Opt. Lett.* **5** 392–4.
8. Dandridge A., Tveten A. B, Kersey A. D and Yurek A. M. 1987 Multiplexing of interferometric sensors using phase generated carrier techniques *J. Lightwave Technol.* **5** 947–52.
9. Brooks J. L., Moslehi B., Kim B. Y. and Shaw H. J. 1987 Time domain addressing of remote fiber optic interferometric sensor arrays *IEEE J. Lightwave Technol.* **5** 1014–23.
10. Henning M. L., Thornton S. W., Carpenter R., Stewart W. J., Dakin J. P. and Wade C. A. 1983 Optical fibre hydrophones with down lead insensitivity *Proc. First Int. Conf. Optical Fiber Sensors. (London).* **1** 23–7.
11. Henning M. L. and Lamb C. 1988 At-sea deployment of a multiplexed fiber optic hydrophone array *Proc. Optical Fibre Sensors Conf. Opt. Soc. Am. (Washington, DC)* **12** 84–91.
12. Dandridge A. 1994 The development of fiber optic sensor systems *Proc. Optical Fibre Sensors Conf. 10 SPIE* **2360** 154–61.
13. Davis A. R., Kirkendall C. K., Dandridge A. and Kersey A. D. 1997 64 channel all optical deployable acoustic array *Proc. Optical Fibre Sensors Conf. Optical Soc. Am. (Washington, DC)* **12** 616–19.
14. Krakenes K. and Bløtekjær K. 1989 Sagnac interferometer for underwater sound detection: noise properties *Opt. Lett.* **14** 1152–4.
15. Knudsen S. 1996 Fiber-optic acoustic sensors based on the Michelson and Sagnac interferometers: responsivity and noise properties *PhD Thesis* ch 2.
16. Udd E. 1983 Fiber-optic acoustic sensor based on the Sagnac interferometer *Proc. SPIE.* **425** 90–5.
17. Vakoc B. J., Digonnet M. J. F. and Kino G. S. 1999. A novel fiber optic sensor array based on the Sagnac interferometer *Proc. SPIE Fiber Optic Sensor Technology and Applications* **3860** 276–84.
18. Dandridge A. and Goldberg L. 1982 Current-induced frequency modulation in diode lasers *Electron. Lett.* **18** 302–4.
19. Kersey A. D., Williams K. J. and Dandridge A. 1989 Characterization of a diode laser pumped Nd:YAG ring laser for fibre sensor applications *Proc. Optical Fibre Sensors Conf. (Firenze) (Springer Proc. in Physics 4)* **9** 172–8.
20. Dandridge A., Tveten A. B., Miles R. O. and Giallorenzi T. G. 1980 Laser noise in fiber-optic interferometer systems *Appl. Phys. Lett.* **37** 526–8.
21. Kane T. J. 1990 Intensity noise in diode-pumped single-frequency Nd–YAG lasers and its control by electronic feedback *IEEE Photon. Technol. Lett.* **2** 244–5.
22. Dagenais D. M., Koo K. P. and Dandridge A. 1991 Demonstration of low-frequency intensity noise reduction for fiber sensors powered by diode-pumped Nd–YAG lasers *IEEE Photon. Technol. Lett.* **4** 519–20.
23. Dandridge A. 1981 Noise reduction in fiber-optic interferometer systems *Appl. Opt.* **20**.
24. Nirvana auto-balanced photoreceiver manual, model 2007 & 2017, New Focus Inc, US.
25. Dandridge A., Tveten A. B., Miles R. O., Jackson D. A. and Giallorenzi T. G. 1981 Single-mode diode laser phase noise *Appl. Phys. Lett.* **38** 77–8.
26. Dandridge A. and Tveten A. B. 1981 Electronic phase-noise suppression in diode lasers *Electron. Lett.* **17** 937–8.
27. Stolpner L., Lee S., Li S., Mehnert A., Mols P., Siala S., Bush J. 2008 Low noise planar external cavity laser for interferometric fiber optic sensors *Proc. SPIE 7004, 19th Int. Conf. Optical Fibre Sensors*, 700457 (Also see RIO Planex external cavity laserdiode), www.rio-lasers.com.
28. Kaminow I. P. 1981 Polarization in optical fibers *IEEE J. Quantum Electron.* **17** 15–22.
29. Bucaro J. A., Lagakos N., Cole J. and Giallorenzi T. 1982 Fibre optic acoustic transduction *Phys. Acoust.* **16** 385–457.
30. Budiansky B., Drucker D. C., Kino G. S. and Rice J. R. 1979 Pressure sensitivity of a clad optical fiber *Appl. Opt.* **18** 4085–8.
31. Lobo Ribeiro A. B., Caleya R. F. and Santos J. L. 1995 Progressive ladder network topology combining interferometric and intensity fiber-optic-based sensors *Appl. Opt.* **34** 6481–8.
32. Hughes R. and Jarzynski J. 1980 Static pressure sensitivity amplification in interferometric fiber optic hydrophones *Appl. Opt.* **19** 99–106.
33. Lagakos N., Schnaus E., Cole J., Jarzynski J. and Bucaro J. 1982 Optimizing fiber coatings for interferometric acoustic sensors *IEEE J. Quantum Electron.* **18** 683–9.
34. Lagakos N. and Bucaro J. A. 1993 Linearly configured embedded fiber-optic acoustic sensor *J. Lightwave Technol.* **11** 639–42.
35. Nash P. J. and Keen J. 1990 Design and construction of practical optical fibre hydrophones *Proc. Inst. Acoust.* **12** 201–12.
36. Knusden S., Havsgard G. B., Christensen Ø., Wang G., Tveten A. and Dandridge A. 1997 Bandwidth limitations due to mechanical resonances of fiber-optic air backed mandrel hydrophones *Proc. Optical Fibre Sensors Conf. Opt. Soc. America (Washington, DC).* **12** 544–7.
37. Wang C. C., Dandridge A., Tveten A. B. and Yurek A. 1994 Very high responsivity fiber optic hydrophones for commercial applications *Proc. Optical Fibre Sensors Conf. 10 SPIE.* **2360** 360–3.
38. Knudsen S. 1996 Fiber-optic acoustic sensors based on the Michelson and Sagnac interferometers: responsivity and noise properties *PhD Thesis* ch 6.
39. Nash P. J., Gallaher A. B. and Hardie D. J. W. 1995 Finite element modelling of optical fibre hydrophones *Proc. Inst. Acoust.* **17** 164–72.
40. Cheng L. K. and De Bruijn D. 1996 Fieldtest of a fibre optic hydrophone *Proc. Optical Fibre Sensors Conf. Advanced Sensing Photonics Japan Soc. Appl. Phys.* **11** 184–7.
41. Sato R., Ishii H., Dobashi K., Kamata H. and Saito S. 1993 Pressure balancing structure for fiber-optic flexural disk acoustic sensor *Japan. J. Appl. Phys.* **32** 2473–6.

42. Kinsley L. E., Frey A. R., Coppens A. B. and Sanders J. V. 1982 *Fundamentals of Acoustics* 3rd edn (New York: Wiley).
43. Waagaard O. H., Havsgard G. B. and Wang G. 2001 An investigation of pressure to acceleration responsivity of fiber-optic hydrophones *J. Light. Technol.* **19** 994–1003.
44. Harris A. J and Castle P. F. 1986 Bend loss measurements on high numerical aperture single-mode fibers as a function of wavelength and bend radius *J. Lightwave Technol.* **4** 34–40.
45. Marcuse D. 1976 Curvature loss formula for optical fibres *J. Opt. Soc. Am.* **6** 216–20.
46. Lagakos L., Ehrenfeuchter P., Hickman T. R., Tveten A. and Bucaro J. A. 1988 Planar flexible fiber-optic interferometric acoustic sensors *Opt. Lett.* **13** 1298–303.
47. Brown D. A., Hofler T. and Garrett S. L. 1989 High-sensitivity, fiber-optic, flexural disk hydrophone with reduced acceleration response *Fiber Integrated Opt.* **8** 169–91.
48. Danielson D. A. and Garrett S. L. 1989 Fibre-optic ellipsoidal flextensional hydrophones *IEEE J. Lightwave Technol.* **7** 1995–2002.
49. Othonos A., Kalli K. 1999 “Fiber Bragg Gratings” Artech House, chap. 4.
50. Lagakos N., Jarzynski J., Cole J. H. and Bucaro J. A. 1986 Frequency and temperature dependence of elastic moduli of polymers *J. Appl. Phys. Lett.* **59** 4017–31.
51. Jackson D. A., Kersey A. K., Corke M. and Jones J. D. C. 1982 Pseudoheterodyne detection scheme for optical interferometers *Electron. Lett.* **18** 1081–3.
52. Kersey A. D., Jackson D. A. and Corke M. 1982 Demodulation scheme fibre interferometric sensors employing laser frequency switching *Electron. Lett.* **19** 102–3.
53. Koo K. P., Tveten A. B., and Dandridge A. 1982 Passive stabilization scheme for fiber interferometers using (3*3) fiber directional couplers *Appl. Phys. Lett.* **41** 616–18.
54. Cole J. H., Danver B. A. and Bucaro J. A. 1982 Synthetic-heterodyne interferometric demodulation *J. Lightwave Technol.* **18** 694–7.
55. Dandridge A., Tveten A. B. and Giallorenzi T. G. 1982 Homodyne demodulation scheme for fiber optic sensors using phase generated carrier *IEEE J. Quantum Electron.* **18** 1647–53.
56. Dandridge A. and Kersey A D 1988 Overview of Mach–Zender sensor technology and applications *Proc. SPIE Fiber Optic and Laser Sensors VI* **985** 34–52.
57. Farshad M. 1994 *Stability of Structures* (New York: Elsevier).
58. Stremler F. G. 1982 *Introduction to Communication Systems* 2nd edn (Reading, MA: Addison-Wesley) section 6.2.
59. Yariv A. 1991 *Optical Electronics* 4th edn (Philadelphia, PA: Saunders) pp. 430–1.
60. Davis C. C. 1996 *Lasers and Electro-Optics Fundamentals and Engineering* 1st edn (Cambridge: Cambridge University Press) pp. 567–8.
61. Wanser K. H. 1994 Theory on thermal phase noise in Michelson and Sagnac fiber interferometers *Proc. SPIE Optical Fibre Sensors Conf. 10* **2360** 584–7.
62. Wanser K. H., Kersey A. D. and Dandridge A. 1993 Measurement of fundamental thermal phase fluctuations in optical fiber *Proc. Optical Fibre Sensors Conf. (Florence) Associazione Elettrotecnica* **9** 255–8.
63. Martinelli M. 1994 Time-reversal of the polarization state in optical fiber circuits *Proc. Optical Fibre Sensors Conf. 10 SPIE* **2360** 312–18
64. Henning M 1985 Improvements in reflectometric fiber optic hydrophones *Proc. 2nd Int. Symp. Optical and Electro Optical Science and Eng. (SPIE), Cannes.*
65. Kersey A. D. and Berkoff T. A. 1990 Novel passive phase noise cancelling technique for interferometric fibre optic sensors *Electron. Lett.* **26** 640–1.
66. Agrawal G. P. 1997 *Fiber-Optic Communication Systems* 2nd edn (New York: Wiley–Interscience) pp. 404–6.
67. Stremler F. G. 1982 *Introduction to Communication Systems* 2nd edn (Reading, MA: Addison-Wesley) p. 293.
68. Kersey A. D. 1990 Multiplexed interferometric fiber sensors *Proc. Optical Fibre Sensors* **7** 313–19.
69. Kersey A. D. and Dandridge A. 1989 Comparative analysis of multiplexing techniques for interferometic fiber sensing *Proc. SPIE Fiber Optic and Laser Sensors* **1120** 236–46.
70. Butter C. D. and Hocker G. B. 1978 Fiber optic strain gauge *Appl. Opt.* **17** 2867–9.
71. Kersey A. D., Dandridge A., Davis A. R., Kirkendall C. K., Marrone M. J. and Gross D. G. 1996 64-element time-division multiplexed interferometric sensor array with EDFA telemetry *Optical Fiber Communications Tech. Digest Series (IEEE Cat. no. 96CH35901) Optical Soc. Am. (Washington, DC)* 270–1.
72. Nash P. J. and Cranch G. A. 1999 Multi-channel optical hydrophone array with time and wavelength division multiplexing *Proc. Optical Fibre Sensors 13 (Kyongju, Korea) SPIE* **3746** 304–7.
73. Dakin J. P., Wade C. A. and Henning M. L. 1984 Novel optical fibre hydrophone array using a single laser source and detector *Electron. Lett.* **20** 53–4.
74. Kersey A. D., Dorsey K. L. and Dandridge A. 1989 Cross talk in a fiber-optic Fabry–Pérot sensor array with ring reflectors *Opt. Lett.* **14** 93–5.
75. Cranch G. A. and Nash P. J. 2001 Large-scale multiplexing of interferometric fibre optic sensors using TDM and DWDM *J. Light. Technol.* **19** 687–99.
76. Vohra S. and Dandridge A. 1996 An hybrid WDM/TDM reflectometric array *Proc. Optical Fiber Sensors Conf. Advanced Sensing Photonics Japan. Soc. Appl. Phys. (Tokyo)* **11** Th 3–29.
77. Kersey A. D., Dandridge A. and Tveten A. B. 1987 Time division multiplexing of interferometric fibre sensors using passive phase generated carrier interrogation *Opt. Lett.* **12** 775–7.
78. Kersey A. D. and Dandridge A. 1989 Ten-element time-division multiplexed interferometric fiber sensor array *Proc. Optical Fibre Sensors Conf.* vol 6 *(Berlin: Springer)* 486–90.
79. Sather J. and Bløtekjær K. 1996 Optical amplifiers in multiplexed sensor systems—theoretical prediction of noise performance *Proc. Optical Fibre Sensors Conf. Advanced Sensing Photonics Japan. Soc. Appl. Phys. (Tokyo)* **11** 518–21.

80. Sæther J. and Bløtekær K. 1997 Optical amplifiers in time domain multiplexed sensor systems *Proc. Optical Fibre Sensors Conf. (OSA Tech. Digest Series 16)* **12** 586–9.
81. Wagener J. L., Hodgson C. W., Digonnet M. J. F. and Shaw H. J. 1997 Novel fiber sensor arrays using erbium-doped fiber amplifiers *J. Lightwave Technol.* **15** 1681–8.
82. Hodgson C. W., Digonnet M. J. F. and Shaw H. J. 1997 Large-scale interferometric fiber sensor arrays with multiple optical amplifiers *Opt. Lett.* **22** 1651–3.
83. Cranch, G. A., Nash, P. J. and Kirkendall, C. K. 2003 Large-scale remotely interrogated arrays of fiber-optic interferometric sensors for underwater acoustic applications *IEEE Sens. J.* **3**(1) 19–30.
84. Marrone M. J, Kersey A. D and Dandridge A. 1992 Polarization independent array configuration based on Michelson interferometer networks *Proc. SPIE Distributed and Multiplexed Fiber Optic Sensors II* **1797** 196–200.
85. Kersey A. D., Marrone M. J. and Dandridge A. 1988 Optimization and stabilization of visibility in interferometric fiber-optic sensors using input-polarization control *J. Lightwave Technol.* **6** 1599–609.
86. Kersey A. K., Dandridge A. and Marrone M. J. 1987 Single-mode fiber pseudo-depolarizer *Proc. SPIE Fiber Optics and Laser Sensors V* **838** 360–4.
87. Ahn J. T. and Kim B. Y. 1995 Fiber-optic sensor array without polarization signal fading *Opt. Lett.* **20** 416–18.
88. Simon A. and Ulrich R. 1977 Evolution of polarization along a single-mode fiber *Appl. Phys. Lett.* **31** 517–20.
89. Frigo N. J., Dandridge A. and Tveten A. B. 1984 Technique for elimination of polarization fading in fibre interferometers *Electron. Lett.* **20** 319–20.
90. Wanser K H and Safar N H 1997 Remote polarization control for fiber-optic interferometers *Opt. Lett.* **12** 217–19.
91. Cranch, G. A., Flockhart, G. M. H. and Kirkendall, C. K. 2008 Distributed feedback fiber laser strain sensors *IEEE Sens. J.* **8**(7) 1161–72.
92. Foster, S., Tikhomirov, A. and Milnes, M. 2007 Fundamental thermal noise in distributed feedback fiber lasers *IEEE J. Quant. Electron.* **43**(5) 378–84.
93. Foster, S., Tikhomirov, A. and Van Velzen, J. 2011 Towards a high performance fiber laser hydrophone *J. Lightwave Technol.* **29**(9) 1335–42.
94. Cranch, G. A., Miller, G. A. and Kirkendall, C. K. 2011 Fiber laser sensors: Enabling the next generation of miniaturized, wideband marine sensors *Paper presented at the Proceedings of SPIE - the International Society for Optical Engineering, 8028*.
95. Foster, S., Tikhomirov, A., Harrison, J. and Van Velzen, J. 2015 Demonstration of an advanced fibre laser hydrophone array in gulf st Vincent *Paper presented at the Proceedings of SPIE - the International Society for Optical Engineering, 9634*.
96. Nash, P., Cranch, G. and Hill D. 2000 Large scale multiplexed fibre-optic arrays for geophysical applications *SPIE Industrial Sensing Systems* **4202** 55–65.
97. Liao Y., Austin, E., Nash, P. J., Kingsley, S. A. and Richardson, D. J. 2012 Highly scalable amplified hybrid TDM/WDM array architecture for interferometric fiber-optic sensor systems *J. Lightwave Technol.* **31**(6).
98. Cranch, G. A., Kirkendall, C. K., Daley, K., Motley, S., Bautista, A., Salzano, J., Nash, P. J., Latchem, J. and Crickmore, R. 2003 Large-scale remotely pumped and interrogated fiber-optic interferometric sensor array *IEEE Photonics Technol. Lett.* **15**(11) 1579–81.
99. Cranch, G. A., Crickmore, R., Kirkendall, C. K., Bautista, A., Daley, K., Motley, S., Salzano, J., Latchem, J. and Nash, P. J. 2004 Acoustic performance of a large-aperture, seabed, fiber-optic hydrophone array *J. Acoust. Soc. Am.* **115**(6) 2848–58.
100. Cole, J. H., Sunderman, C., Tveten, A. B., Kirkendall, C., and Dandridge, A. 2002 Preliminary investigation of air-included polymer coatings for enhanced sensitivity of fiber-optic acoustic sensors *Paper presented at the 2002 15th Optical Fiber Sensors Conference Technical Digest, OFS 2002*, 317–20.
101. Cranch, G. A., Miller, G. A. and Kirkendall, C. K. 2012 Fiber-optic, cantilever-type acoustic motion velocity hydrophone *J. Acoust. Soc. Am.*, **132**(1), 103–14.
102. Cranch, G. A., Lane, J. E., Miller, G. A. and Lou, J. W. 2013 Low frequency driven oscillations of cantilevers in viscous fluids at very low Reynolds number *J. Appl. Phys.* 113(19).
103. Gardner, D. L. and Garrett, S. L. 1987 Fiber optic seismic sensor *SPIE Fiber Optic and Laser Sensors V* **838** 271–278.
104. Cranch, G. A. and Nash, P. J. 2000 High-responsivity fiber-optic flexural disk accelerometers *J. Lightwave Technol.* **18**(9) 1233–43. doi:10.1109/50.871700.
105. Nakstad H., Langhammer J. and Eriksrud M. 2011 Fibre optic permanent reservoir monitoring breakthrough *12th Int. Congress of the Brazilian Geophysical Society.*

FURTHER READING

Cranch G. A. 2015 Fiber-optic sensor multiplexing principles *Handbook of Optical Sensors* ed. J. Santos, F. Farahi (BocaRaton, FL: CRC press/Taylor & Francis) ch. 15.

Dandridge A. 1991 Fiber optic sensors based on the Mach–Zender and Michelson interferometers *Fiber Optic Sensors: An Introduction for Engineers and Scientists* ed E. Udd (New York: Wiley) ch 10, pp. 271–323.

Dandridge A and Kirkendall C. 2002 Passive fiber optic sensor networks *Handbook of Optical Fibre Sensing Technology* ed. J. M. Lopez-Higuera (John Wiley & Sons Ltd).

Nash P. J. 1996 Review of interferometric optical fibre hydrophone technology *IEE Proc.-Radar, Sonar Navig.* **143** 204–9.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

7

Laser Stabilization for Precision Measurements

G. P. Barwood and P. Gill

CONTENTS

7.1 Basic Spatial and Spectral Characteristics of Lasers....111
7.2 Advantages of Frequency Stabilization....112
7.3 Applications of Frequency-Stabilized Lasers....112
  7.3.1 Frequency-Stabilized Lasers as Sources for Dimensional Interferometry....112
  7.3.2 Interferometry for Gravitational Wave Detection....114
  7.3.3 The Determination of Fundamental Constants....114
7.4 Gas-Cell-Absorption-Based Stabilization Techniques....115
  7.4.1 Frequency Stabilization of a He-Ne Laser to the Gain Curve....115
  7.4.2 Frequency Stabilization Based on Doppler-Limited Absorption....116
  7.4.3 Frequency-Stabilization-Based Doppler-Free Spectroscopy....116
  7.4.4 Frequency-Stabilized Lasers Referenced to Iodine, Rubidium and Acetylene....118
7.5 Evaluation of Frequency Stability and Reproducibility....121
7.6 Cavity-Stabilization Techniques....122
7.7 Summary....124
References....125
Further Reading....126

7.1 Basic Spatial and Spectral Characteristics of Lasers

Lasers are uniquely suited to applications in precision measurement due to their output beam spatial and spectral characteristics. They emit light in a near diffraction-limited beam and may also emit in a single longitudinal mode, with a narrow linewidth. By way of introduction to this chapter, a few comments covering spatial characteristics are given, in so far as they can affect the stabilization and application of narrow linewidth lasers. Subsequently, we concentrate solely on spectral characteristics.

Ideally, the emitted light from a laser has a Gaussian spatial profile (the TEM_{00} mode), although more complicated spatial mode patterns are possible. The intensity profile is then of the form $I = I_0 \exp(-2r^2/w_0^2)$, where r is the distance from the beam centre and w_0 is the $1/e$ amplitude radius of the beam [1]. He-Ne and YAG lasers are examples of systems that emit near-Gaussian beams. However, optical elements in the cavity, for example, Brewster windows or even the gain medium itself, can make the beam slightly astigmatic and, therefore, non-Gaussian. Diode lasers are a common example of a system with a poor spatial beam profile [2]. In this case, the beam profile is determined by the shape of the semiconductor junction. Since the junction's cross section tends to be rectangular, even for low-power lasers, the far-field beam pattern is elliptical. Special beam-shaping optics are often used in order to achieve a more circular Gaussian beam profile. Near-Gaussian beam quality is important for frequency stabilization to transitions in gas or vapour cells, where higher-order spatial modes can give rise to small frequency shifts. It is also important during frequency-stabilizing lasers to optical cavities; otherwise, the input coupling efficiency will be low. Poor spatial-mode quality can also limit the use of lasers in interferometry, causing beam distortion when propagating over long distances.

Spectrally, lasers may emit light in either a single longitudinal mode or via several modes throughout their spectral gain profile. Some applications in precision measurement, for example, displacement-measuring interferometry, can use either single- or multi-mode laser sources. The latter can only be used where the optical path difference is less than a few tens of centimetres. Many other applications in spectroscopy and optical frequency metrology inevitably require the use of a single-mode source. There is a variety of methods available for forcing multi-mode lasers to emit in only a single mode. A common technique is to use a low-finesse etalon inside the cavity, creating a small loss at frequencies other than at the desired mode. Competition between adjacent modes in the laser then ensures that only this mode lases.

With a single longitudinal mode laser, there are a number of factors which can determine the spectral linewidth. In many common cw single longitudinal mode lasers, such as He-Ne devices, the linewidth is often determined by the level of air- or ground-borne vibrations in the laboratory, perturbing the cavity length and, hence, the emitted frequency. If these vibration levels are reduced, then in He-Ne lasers, there remains the problem of noise due to the discharge. In optically pumped solid-state lasers, where there is no discharge, the laser linewidth is far smaller. In particular, the diode-pumped Nd:YAG laser at 1064 nm can have a linewidth in the region of 10 kHz. The frequency-doubled Nd:YAG laser at 532 nm is thus well suited to frequency stabilization, since the narrow linewidth should ensure good short-term frequency stability. Laser diodes, which have a much larger tuning range (several THz for a given device), however, have a frequency noise problem associated with their short cavity length. In this case, the dominant problem is with the very short, low-finesse cavity of a diode. This dominant noise source arises from noise due to spontaneous emission in the cavity. The linewidth of a diode laser is further increased because, following each spontaneous emission event, there is a small change $\Delta n''$ in the imaginary part of the refractive index of the semiconductor. This is caused by a change in the carrier density, which also changes the real part of the refractive index by $\Delta n'$. Defining $\alpha = \Delta n'/\Delta n''$, the formula for the full-width-at-half-maximum (FWHM) linewidth, γ, is

$$\gamma = \frac{2\pi h f \eta \Delta f_{1/2}^2 (1+\alpha^2)}{P} \tag{7.1}$$

This is the modified Schawlow–Townes formula: the original formula has $\alpha = 0$ but, for semiconductor lasers, $\alpha \simeq 5$ is more appropriate [3,4]. Equation (7.1) shows that the FWHM depends on the cavity half-width half maximum ($\Delta f_{1/2}$) and the emitted power (P). In equation (7.1), h is the Planck constant, f is the optical frequency and η is the inversion factor ($\eta = N_2/(N_2 - N_1)$), with N denoting the occupancy of the two levels. For a diode laser, with a cavity length of 1 mm and a mirror reflectivity of 30%, the limit is in the region of a few tens of MHz, and therefore, this dominates over other effects. In contrast, for a typical He-Ne laser, this Schawlow–Townes limit is in the region of a few mHz and so other effects, principally acoustic vibration or He-Ne discharge noise, always dominate.

7.2 Advantages of Frequency Stabilization

For many applications in precision measurements, it is important to have good laser frequency stability and to know the absolute optical frequency. Knowledge of the absolute frequency at a particular level of accuracy implies that the laser also needs to be reproducible at that level from day to day. In order to develop a frequency-stabilized laser, it is generally necessary to use a single-mode system. Common choices of single-mode laser include the He-Ne, Nd:YAG, diode or titanium–sapphire (Ti:S) laser. The Ti:S laser has now almost completely superseded the dye laser [5], where a tunable high-power source is required. These lasers have been listed in order for increasing tuning range. He-Ne lasers, for example, emit only within a Doppler-broadened gain profile of typically 1 GHz, at very specific wavelengths determined by the neon transitions. In the visible part of the spectrum, these wavelengths are 543, 594, 612, 633 and 640 nm. The optically pumped Nd-doped YAG tunes over a few tens of GHz. The output wavelength is determined by transitions in Nd, the most common being at 1064 nm. This output is often doubled to produce 532 nm and higher harmonics are possible. Laser diodes can be produced to cover most wavelengths between 630 and 1000 nm, with other wavelengths available at 400, 1300 and 1500–1600 nm. Individual units can tune several THz. However, in order to develop a stabilized laser at a particular frequency, the most difficult problem is generally to source a diode which can tune to exactly the desired frequency. Manufacturers may be able to provide this service, and some companies hold considerable stock to allow customer selection. Finally, a Ti:S laser can be used to cover the entire region between 700 and 1000 nm. The high output power also allows frequency doubling from 350 to 500 nm. All these lasers can be designed to emit in a single longitudinal mode, but as can readily be appreciated, this single mode can emit anywhere within a considerable range. For many applications it is advantageous to restrict laser operation within a much tighter bandwidth by frequency stabilization.

7.3 Applications of Frequency-Stabilized Lasers

Stabilized lasers can be used in a number of applications, such as interferometry (including dimensional metrology and gravitational wave detection), measurements of fundamental constants, including investigations regarding their possible variation with time, frequency standards and optical clocks. In the following Sections (7.3.1–7.3.3), some of these applications are discussed in outline.

7.3.1 Frequency-Stabilized Lasers as Sources for Dimensional Interferometry

Interferometry is a widely used technique for the precision measurement of length. One of the most common interferometer systems for length (or, more strictly, displacement) measurement is the Michelson interferometer, shown in Figure 7.1 (e.g. Ref. [6]). In this system, light from the laser is split, with one beam going to a fixed (or reference) reflector (usually a corner cube) and the other beam travelling along the path to be measured. On return, the beams are re-combined at the beamsplitter and produce optical fringes. The resulting intensity is given by

$$I = \left| E_{\text{meas}} \exp(\mathrm{i}(wt + 2kL)) + E_{\text{ref}} \exp(\mathrm{i}\, wt) \right|^2 \tag{7.2}$$

where L is the distance between the beamsplitter and the measurement corner cube. The reference path and measurement path electric field amplitudes at the detector are denoted by E_{ref} and E_{meas}, respectively. It is easily shown that this equation becomes

$$I = E_{\text{meas}}^2 + E_{\text{ref}}^2 + 2E_{\text{ref}} E_{\text{meas}} \cos 2kL. \tag{7.3}$$

The intensity function, therefore, has a periodicity given by $kL = n\pi$, where n is an integer and the wavenumber k is $2\pi/\lambda$. The detected periodicity or ‘fringe’ can be electronically

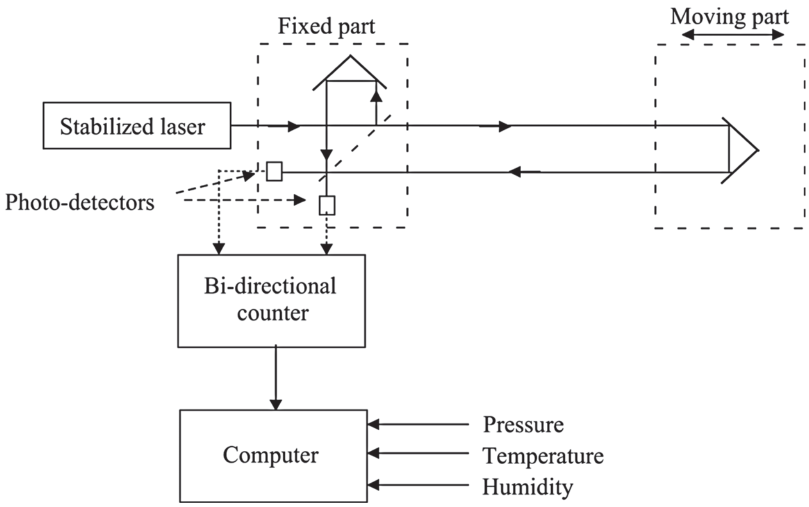


FIGURE 7.1 Michelson interferometer system measuring displacement.

counted as a function of displacement of the measurement retro-reflector, using the initial position of the retro-reflector as the zero reference position. The displacement is then given by $L = \frac{1}{2}n\lambda$, where n is the number of fringes counted and the wavelength is λ. For a 1 m displacement, for example, this corresponds to around three million fringes. Provided we know the wavelength very well, then this gives a simple but accurate means to determine displacement. However, we should note that if this measurement is performed in air, then the air refractive index must also be taken into account [7–9], so that a knowledge of the laser vacuum wavelength (and, hence, the frequency) is only part of the measurement problem. In most systems, the air refractive index is calculated from the air temperature, pressure and relative humidity. System measurement capabilities of 1 part in 10^6 over the range of 1 m to tens of metres are typical for systems where the refractive index correction is automatically compensated for. This capability can be improved to approximately 1 part in 10^7, provided more accurate environmental measurements are made, and the laser is stabilized to a few parts in 10^8 or better. Alternatively, interferometric techniques may also be used either to measure the refractive index absolutely [10] or track changes in the refractive index. Although we require a known and stable laser frequency in order to obtain the most accurate measurements, a multi-mode He-Ne laser may be used if we only require an accuracy of no better than a few in 10^5. This is because the gain profile of the laser is only around 1 GHz (two parts in 10^6 of the optical frequency).

From equation (7.3), the intensity output of the Michelson interferometer varies between $I_{\max}=(E_{\text{meas}}+E_{\text{ref}})^2$ and $I_{\min}=(E_{\text{meas}}-E_{\text{ref}})^2$ and we may define a fringe visibility function as follows:

$$V = \frac{(I_{\max} - I_{\min})}{(I_{\max} + I_{\min})}. \tag{7.4}$$

In this case, considering only the effect of unequal beam intensities in the two arms of the interferometer, we have

$$V = \frac{2E_{\text{ref}}E_{\text{meas}}}{(E_{\text{ref}}^2 + E_{\text{meas}}^2)}. \tag{7.5}$$

The fringe visibility is, therefore, unity in the ideal case where the measurement and reference amplitudes are equal. However, there are other factors which can reduce this visibility from unity and the most important of these is probably the laser linewidth. It can be shown that the fringe visibility is proportional to the Fourier transform of the laser spectral profile. For example, for a laser diode with a Lorentzian profile and FWHM of $c/\pi L_c$, the fringe visibility varies as $V\,\alpha\,\exp(-L/L_c)$, where L_c is the laser coherence length. This demonstrates another reason for frequency stabilizing the laser, namely, to prevent the laser from tuning to a region where the spectrum is multi-mode with a corresponding reduction in fringe visibility. This is particularly the case with He-Ne lasers. With a narrow-linewidth source (for example, <1 MHz), the distance that can be measured is generally determined by air turbulence, which causes rapid changes in the fringe visibility as the return-beam position varies spatially. For long distances, the optical beam must also be fairly large (typically 10 mm in diameter) in order to prevent too much beam divergence over the return distance. Commercial systems are available that can measure displacements of up to 80 m.

It is important to emphasize that the Michelson system described in the previous paragraph is measuring the displacement of the corner cube and not an absolute distance. For distance measurement, other techniques are available—for example, swept frequency interferometry [11,12 also references therein], where the interferometer path-lengths are kept fixed but the laser frequency is scanned over a known frequency range in order to obtain scanning fringes. This will require the frequency at both ends of the scan to be accurately known.

In order to calibrate material length standard artefacts, such as gauge blocks, the method of excess fringe fractions is used (for a diagram showing a gauge block interferometer system, see Figure 7.1). With this method, one end of the gauge block is wrung to a polished base-plate and wedged fringes are formed between the base-plate and a reference flat, and between the top end of the gauge block and the reference flat.[1] For a single fixed wavelength, a fringe discontinuity (i.e. fringe fraction) between the two sets of wedged fringes is observed. However, this does not provide information on the difference in order number corresponding to the length of the gauge. In order to determine this, fringe fractions are measured at a number of wavelengths and, together with a nominal value for gauge length determined by initial lower accuracy methods, the order number can be unambiguously assigned and a higher accuracy measurement of length achieved. Length measurement accuracies of approximately one part in 10^7 are possible using this technique, limited by environmental and surface features, rather than the laser frequency uncertainty.

7.3.2 Interferometry for Gravitational Wave Detection

Gravitational waves, originally predicted by Einstein, are estimated to produce distortions in space, although at levels well below that which are achieved by conventional interferometry [13]. For example, violent astrophysical events, such as supernova explosions, require strain sensitivity measurements of around a part in 10^{21}. Although this represents a major difficulty, it is only necessary to achieve this in a certain bandwidth. For ground-based interferometers, this is typically a few tens of Hz in the hundreds of Hz region. A number of novel 'recycling' interferometer systems have been developed for this purpose and shown to have sensitivities close to this required value. For example, in the power recycling mode [14], light returning towards the laser is reflected back into the interferometer in order to assist the power build-up. It can be shown [13] that for powers in the region of 7×10^7 W optical sensitivity close to the standard quantum limit is achieved. For more details of specific systems, see the websites cited in the footnotes. There are a number of groups worldwide with experiments targeted on gravity wave detection and developing sensitive ground-based interferometers for this purpose. Interferometer arm dimensions are typically 600 m to 3 km, with evacuated tubes and interferometer optics and mirrors suspended from seismically damped supports. The UK–German ground-based interferometer GEO600, based around a 600 m interferometer system, is currently being tested near Hanover. This is a recycling interferometer system, based on a 10 W single-mode YAG laser at 1064 nm, and has a sensitivity of 10^{-20}–10^{-21}. Power recycling increases the power in the interferometer to 10 kW. It is designed to operate in a 60-Hz bandwidth between 50 Hz and 1.5 kHz. Space-based interferometers offer improved resolution and can detect gravity waves at much lower frequencies. Other gravitational wave detector facilities include, for example, VIRGO (France–Italy), TAMA (Japan) and LIGO (USA).[2] There is a proposal for a future space-based interferometer (LISA), with path lengths of around 5×10^9 m. This is around 1/30 of the distance between the earth and the sun.

7.3.3 The Determination of Fundamental Constants

Frequency-stabilized lasers also have a key role to play in the determination of some fundamental constants, through their use as references for tunable sources in high-resolution spectroscopy. Rydberg originally observed that the wavelengths of known lines in hydrogen fitted a simple pattern. As is now well known, these lines are, at vacuum wavelengths λ_0, given by

$$\frac{1}{\lambda} = R\left(\frac{1}{p^2} - \frac{1}{q^2}\right) \tag{7.6}$$

where p and q are integers and R is the Rydberg constant. The Bohr theory of the hydrogen atom allows the determination of R:

$$R_\infty = \frac{m_e e^4}{8\varepsilon_0^2 h^3 c} \tag{7.7}$$

where m_e and e are the electron mass and charge, ε_0 is the permittivity of free space, h is the Planck constant and c is the speed of light. The subscript used for R indicates that the Bohr model assumes that the proton mass is infinite. The finite proton mass is just one of a large number of corrections that need applying to the simple formula to explain various frequency shifts and splittings that are observed in hydrogen. There has been a considerable activity in the precision measurements of frequencies of transitions both in hydrogen and hydrogen-like species (i.e. deuterium, positronium and muonium). Local frequency standards were developed based on tellurium, in order to provide local high-accuracy references for hydrogenic work [15]. Within a few years, measurements of hydrogenic frequency standards had progressed to the point where the narrow 1S–2S two-photon transition at 243 nm had been observed and measured to a precision of two parts in 10^{14} [16]. High-accuracy comparison of measured and calculated hydrogen-like frequencies now provide stringent tests of theories such as quantum electrodynamics. More details of this subject can be found in [17,18].

The long-term reproducibility of some high-accuracy lasers and frequency standards can also be used to investigate the constancy of the spectroscopic fine structure constant, $\alpha = (\mu_0 c e^2/2h)$. Since this constant is dimensionless, any change is independent of the units in which it is expressed. Both laboratory-based and astrophysical observations have been published: astrophysical data appear to indicate a real change as a function of the redshift [19], although earlier measurements failed to detect any significant change (see, for example, Ref. [20]). Laboratory experiments (see, for example, Ref. [21]) have so far been unable to establish any evidence for a change with time. These experiments rely on different standards, whether optical or microwave, having different α dependences, so that measurements of ratios between them will vary. In the optical case, the availability of femtosecond laser comb technology to measure optical frequency ratios to high accuracy between very stable lasers and optical frequency standards should help make significant advances in this area. For more details on fundamental constants, a review has been published [18] and a brief introduction to measurements on variations in α is discussed by Ref. [22].

---

[1] For details of a commercially available gauge block interferometer system, see http://www.npl.co.uk/length/dmet/services/equipment/gbi.html.

[2] More information on gravitational waves is available at http:www.physics.gla.ac.uk/gwg/index.html (Institute for Gravitational Research, Glasgow); http://www.gro600.uni-hannover.de/ for the GEO600 project in Hanover, http://www.ligo.caltech.edu/ for the US Ligo projects; http://www.virgo.infn.it/ for the joint French-Italian VIRGO project; http//tamago.mtk.nao.ac.jp/ for 'Tama300' in Japan; http://lisa.jpl.nasa.goc/ for 'LISA' and http://www.anu.edu.au/Physics/ACIGA/ for a project based in Australia.

7.4 Gas-Cell-Absorption-Based Stabilization Techniques

This section describes some of the more common techniques available for frequency-stabilizing lasers, using transitions in gases, vapours or discharges. Typically, these transitions have linewidths of a few MHz but the observed feature width is broadened due to the movement of the atoms and the resulting Doppler shift. Since the velocity distribution of the atoms is Gaussian, the transition profile of the Doppler-broadened feature will also be Gaussian and typically has a linewidth of a few hundred MHz. Where the spectral widths are limited in this way, the resolution is said to be Doppler-limited. Fortunately, techniques are available to observe the transitions without the Doppler broadening, giving Doppler-free resolution, and these methods will be discussed in Sections 4.4.3 and 4.4.4.

Initially, we discuss Doppler-limited methods of laser frequency stabilization, as applied in the case of He-Ne and diode lasers. In the case of the He-Ne laser, the Doppler-broadened gain profile is around 1 GHz wide and techniques are available to frequency-stabilize the laser to the centre of the gain profile. These provide an easy-to-use system with reproducibilities in the range of one part in 10^8 to one part in 10^7. These methods are so useful and straightforward that systems are commercially available.

7.4.1 Frequency Stabilization of a He-Ne Laser to the Gain Curve

One of the simplest and most widely used methods of frequency stabilization is to use a He-Ne laser with a gain tube of around 30 cm in length [23,24]. After turning on, the laser plasma tube gradually heats up and the cavity expands. The frequency of the nth mode of the cavity is given by $f_n = nc/2L$, for a cavity of length L. As the cavity expands, the frequencies of individual modes therefore decrease, drifting through the gain profile. Typically, for a 30-cm cavity length, the gain profile will support two or three modes. The three-mode case occurs with one dominant mode close to the gain centre, whereas the two-mode case occurs with both modes nominally equidistant from the gain centre. Due to mode competition, each mode is orthogonally polarized to the adjacent one. As a result, for the two-mode case, the modes can be split with a polarizing beamsplitter, as shown in Figure 7.2. Two detectors are then used to measure the power in these two modes and the difference in power can be used to generate an error signal. This difference signal can then be integrated and fed back onto heaters which control the cavity length. Integration of the error signal ensures that, at low frequencies, the servo system gain corresponds to the maximum open-loop gain of the amplifier. This minimizes the error signal, which will inevitably be present, if the laser cavity length is slowly drifting. An output polarizer is used to select one of the two modes as the main output beam. Relatively straightforward circuitry allows the design of a stabilized laser system which turns on from cold and automatically frequency-stabilizes in this way, generally in around 20 min. This is the basis of a number of 633-nm frequency-stabilized He-Ne lasers, currently on the market. A slightly more elaborate algorithm is employed for lasers at the lower gain lines at 543 and 612 nm, which generally emit in three longitudinal modes at the centre of the gain profile [25].

This two-mode balanced stabilization technique generates an error signal based on the Doppler-broadened gain profile of the He-Ne laser. Two further schemes exist which make use of the Doppler-free neon feature at the centre of the gain profile.

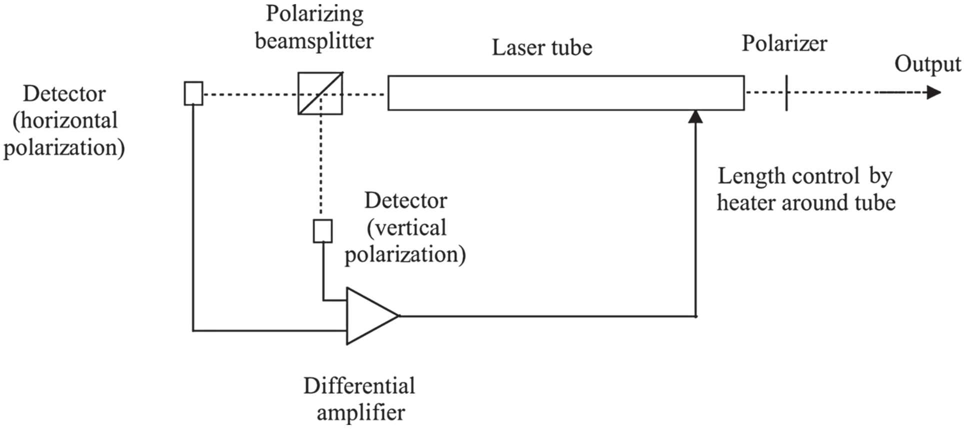

Polarizing beamsplitter
Laser tube
Polarizer
Output
Detector (horizontal polarization)
Length control by heater around tube
Detector (vertical polarization)
Differential amplifier

FIGURE 7.2 Schematic diagram for the frequency stabilization of a two-mode He-Ne laser by balancing the two-mode intensities.

The first of these is based on the Zeeman effect, and the method requires a laser tube sufficiently short that it lases on only one longitudinal mode at the centre of the gain profile. The tube should also have no polarization-selective optics, such as a Brewster-angled window. The He-Ne discharge is placed in an axial magnetic field [26,27], which splits the gain curve into two opposite circularly polarized components, σ_+ and σ_-. Both these Zeeman components provide gain but their cavity mode frequencies are non-degenerate on account of their different refractive index dependences on the magnetic field. In low fields, the frequency difference is so small that the frequencies of the two modes usually lock together. However, in a magnetic field of 10 mT, the refractivity difference becomes sufficiently different that the two modes separate in frequency and a Zeeman beat of a few hundred kHz between the modes is observed. The difference in this refractivity changes as the laser is tuned through the central portion of the gain curve, resulting in a Zeeman beat maximum at line centre. This detected Zeeman beat frequency signal is used to stabilize the laser.

An early frequency stabilization technique, using the Lamb dip, made use of the gain competition experienced by the two travelling-wave components of the standing-wave cavity mode. In general, atoms with velocity $+v$ and $-v$ can contribute to the gain experienced by a cavity mode at the frequency corresponding to either of these velocity classes. However, at line centre, only one set of neon atoms (those with a zero longitudinal velocity component) can contribute to the gain, giving rise to a slight dip in the output power of the mode at this point. In order to stabilize to this Lamb dip, the laser cavity length and, hence, the frequency are modulated by means of a piezoelectrically driven cavity mirror, at a frequency of a few hundred Hz. The technique of frequency modulating the laser and subsequent de-modulation of the power change is now so widely used that we should consider it here in some detail. In general, we have a situation where the power varies with frequency, i.e. $P=g(f)$, and we are particularly interested in the behaviour of this function around some maximum or minimum value at $f=f_0$. We then modulate the laser frequency at an angular frequency Ω with amplitude f_m, so that the power output varies with time and frequency de-tuning as

$$P(t) \ = g(f + f_m sin\Omega t) \tag{7.8}$$

Expanding this function to first order in f_m gives:

$$P(t) = g(f) + f_\mathrm{m}\sin(\Omega t)\,g'(f) + \mathrm{O}\left(f_\mathrm{m}^2\right) \tag{7.9}$$

where the prime denotes a derivative and the terminology O( ) indicates the order of higher terms in the expansion. A phase-sensitive detector (or double-balanced mixer (DBM)) will pick out the signal at Ω, which is $f_\mathrm{m}g'(f)$, so giving the first derivative of the spectral profile. This is an ideal feature shape to produce an error signal, since it is zero at the centre of the Lamb dip and is of different sign either side of the centre. The Doppler-broadened gain curve will also give a slowly varying derivative signal, and this is also zero at the centre of the Lamb dip. As we shall see later, for Doppler-free absorption spectroscopy, this is not generally the case, and so we will need to use additional techniques to remove this Doppler-limited background.

7.4.2 Frequency Stabilization Based on Doppler-Limited Absorption

An alternative method of providing frequency stabilization, giving modest levels of stability, is by referencing to the linear absorption of a molecule or atom in a cell. This technique is advantageous in that the frequency reference providing the stabilization discriminant is decoupled from the laser gain medium. As a result, it is much more widely applicable to different types of laser, whereas the gain-related techniques already described are applicable, in the main, to the He-Ne laser. A secondary advantage is that the stabilization signal is derived from a quiescent absorption rather than an excited discharge. In absorption spectroscopy, the laser intensity (I) transmitted through a cell of length L is given by the Beer law, $I=I_0\exp(-\alpha L)$, where I_o is the incident intensity and the absorbance is αL [28]. The absorption coefficient a is given by the product of the molecular number density, N, and the absorption cross section, σ,

$$\alpha = N\sigma \tag{7.10}$$

As the laser frequency f is scanned through a single Doppler-broadened linear absorption, the cross section, σ, varies as

$$\sigma = \frac{S}{\gamma\sqrt{\pi}}\exp\left(-(f - f_0)^2/\gamma^2\right) \tag{7.11}$$

where $2\gamma\sqrt{\ln 2}$ is the FWHM of the absorption and f_0 is the laser frequency at line centre. The line-strength S of the transition is the total cross section integrated over all frequencies. It follows from equation (7.11) that the absorption cross section σ_0 at line centre is given by

$$\sigma_0 = \frac{S}{\gamma\sqrt{\pi}} \tag{7.12}$$

Although linear absorption Doppler-limited spectroscopy is less widely applied than the Doppler-free techniques described in Section 7.4.3, an example of this method is the stabilization of laser diodes at 1500 nm to acetylene and other molecules [29,30]. A frequency-doubled Nd:YAG laser has also been stabilized to linear absorptions in iodine at 532 nm [31] and a stability of 2.3×10^{-8} was reported. Linear absorption spectroscopy was also used to frequency-stabilize a frequency-doubled diode laser system at 422 nm, using Doppler-broadened transitions in rubidium [32].

7.4.3 Frequency-Stabilization-Based Doppler-Free Spectroscopy

In order to achieve better frequency stability and reproducibility from absorptions in molecules or atoms in a cell, we need to use Doppler-free spectroscopy. This produces far narrower features for frequency locking—typically a few MHz or less compared to the 500 MHz observed for Doppler-limited features. When an absorption cell is placed within a laser cavity, the laser gain profile is modified by any Doppler-limited linear absorptions coincident with the lasing wavelength. In addition, a series of weak hyperfine saturated spectra (inverted Lamb dips) can be observed as the laser is tuned. These features arise since the standing-wave cavity mode sees a narrow region of decreased absorption loss when it is within a natural width of the hyperfine absorption line centre. In effect, the cavity mode standing wave comprises two oppositely directed travelling-wave components, each of which interacts with a particular positive or negative velocity component (dependent on the direction) of the atom's or molecule's thermal velocity distribution which gives rise to the Doppler width. As the laser mode is tuned within a natural linewidth of the hyperfine component centre frequency, it interacts only with the zero-velocity part of the distribution and there are reduced numbers of absorbing molecules within this velocity group. As a result, the absorption is reduced within this narrow region and the laser output increases accordingly, resulting in the saturated spectra superimposed on the tuning curve. This is a standard technique for the removal of Doppler broadening due to absorber thermal motion and, together with the use of a low-pressure absorption cell, provides the means for a long-term narrow spectral reference which is relatively perturbation-free and very reproducible with time.

In order to observe these weak saturated features with good signal-to-noise ratios and then frequency stabilize, standard techniques of signal recovery using phase-sensitive detection (PSD) are used. The laser cavity length (frequency) is modulated by a piezoelectric tube (PZT) to a depth of a few MHz at a rate of typically around 500 Hz. The third derivative of the laser-power tuning curve (obtained by PZT tuning of the cavity length) is recovered by PSD at the third harmonic of the modulation frequency. This is readily appreciated from the earlier discussions on observation of the first derivative of the Lamb dip. If the expansion of equation (7.9) is taken further, we obtain

$$P(t) = g(f) + f_{\mathrm{m}} g'(f)\sin\Omega t + \frac{1}{2} f_{\mathrm{m}}^2 g''(f)\sin^2\Omega t + \frac{1}{6} f_{\mathrm{m}}^3 g'''(f)\sin^3\Omega t + O(f_{\mathrm{m}}^4). \tag{7.13}$$

If we now modify our PSD electronics, so that it looks for signals at the third harmonic of the modulation frequency, we see that it will predominantly be observing the third derivative of the profile, since $\sin^3 \Omega t$ itself contains a term in $\sin 3\Omega t$. Although the third derivative does contain terms relating to the Doppler-limited changes in the iodine absorption and neon gain profile, these are very much smaller than observed in the case of first harmonic detection and the offsets introduced are usually negligible. So, this technique provides a zero-crossing discriminant on a flat background free from large offsets and suitable for stabilization purposes.

However, in modulating the laser within the frequency response range of a PZT (typically a few hundreds of Hz to a few kHz), we are modulating the laser at a frequency where much of the frequency and intensity noise is. This noise may be acoustic or arise from plasma noise in the case of a He-Ne laser. With a diode laser, the frequency noise is 1/*f* and so arises predominantly at low frequencies. In many cases, therefore, it is desirable to modulate the laser at frequencies of several MHz or above. This has the added benefit of allowing us to increase the laser servo bandwidth, which has to be less than the modulation frequency. This enables significant reduction of the laser frequency noise. Frequency modulation (FM) spectroscopy was first demonstrated in the 1980s [33–35] and a typical system or the Doppler-free case in iodine is shown in Figure 7.3.

The probe beam through the iodine cell is phase modulated at an angular frequency Ω, so that the output spectrum is

$$E = E_0 \exp(i(\omega t + \beta \sin\Omega t)) \tag{7.14}$$

where β is termed as the modulation index. This expression is mathematically identical to

$$E = E_0 \exp(\mathrm{i}\,\omega t)\sum_n J_n(\beta)\cos n\Omega t \tag{7.15}$$

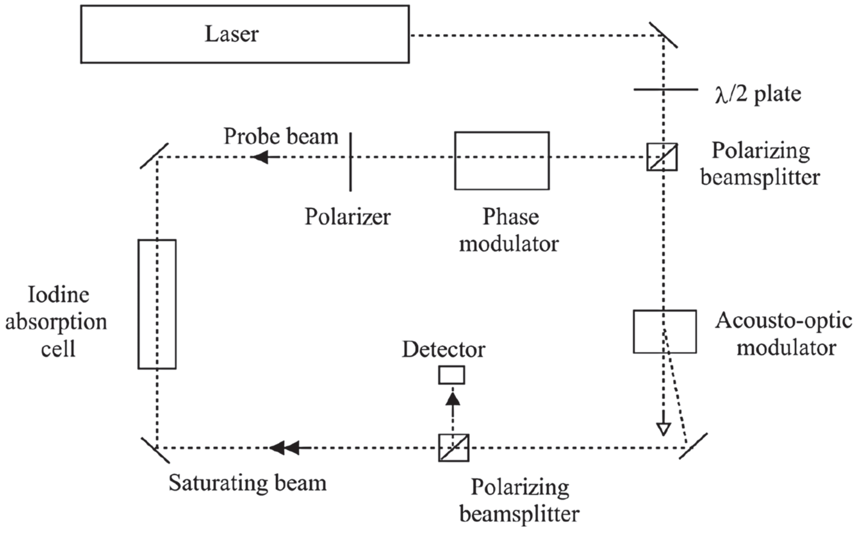


FIGURE 7.3 Experimental arrangement for FM Doppler-free spectroscopy.

so that, provided β is less than 1, this may be approximated as a carrier with a pair of sidebands at frequencies $\pm\Omega/2\pi$. In the absence of any interaction with the absorption cell, equation (7.15) represents pure phase modulation—no amplitude modulation is involved. Each sideband forms a beat with the carrier but the pairs of beats are in anti-phase and so cancel out. In Figure 7.3, the probe and saturating beam overlap in the cell and so the probe beam transmission will be slightly greater as the laser is tuned through the natural Lorentzian width of an iodine transition. The mechanism producing this is nominally the same as that for an inverted Lamb dip using an absorption cell within the laser cavity. As the phase-modulated probe beam is scanned through the transition, the three frequencies (carrier and two sidebands) are absorbed by differing amounts. Therefore, some amplitude modulation is introduced and this is detected and demodulated by a DBM. The profile that is produced is of a characteristic shape, which may be calculated [36] and has a zero cross-over at line centre and so is suitable for frequency-locking. The shape is also dependent upon the relative phases of the reference signal entering the DBM and the phase modulator. Although this method produces FM features suitable for frequency control, it produces them on a slowly varying background which arises from the Doppler-limited absorption. In Figure 7.3, this is removed by amplitude modulation of the saturating beam, which amplitude-modulates the detected saturated signal but not the Doppler-limited background. A phase-sensitive detector, operating at the amplitude chopping frequency, demodulates the DBM output and produces the FM features on a flat background.

7.4.4 Frequency-Stabilized Lasers Referenced to Iodine, Rubidium and Acetylene

The question of which particular atomic or molecular absorber to use depends on the wavelength of the laser. For fixed-frequency lasers, such as He-Ne lasers, operating in the 500–650 nm region, iodine is usually chosen [37]. The electronic transition of the iodine diatomic molecule has associated with it a vast number of molecular rotational and vibrational absorption lines, distributed throughout this visible region, with separations between neighbouring absorptions of a few GHz or less. Each of these lines has an underlying hyperfine structure, which arises from the 5/2 spin of the ^{127}I nucleus. For an even value of J (the rotational quantum number) the total nuclear spin takes the values of 1, 3 or 5 and the transition is, therefore, split into 21 hyperfine levels [38]. For an odd value of J, the total nuclear spin takes the values 0, 2 or 4 and the transition is split into 15 hyperfine levels. There is, therefore, a good probability that any laser that can be tuned across a few hundred MHz can access at least some of these hyperfine states. Iodine vapour can also conveniently be confined in a short silica cell, and its pressure controlled through the use of a cold finger, whilst the vapour temperature can be controlled by heaters around the cell body. At wavelengths longer than 650 nm, the absorptions become weaker as the transitions have to involve ground states with a high vibrational quantum number and low occupancy. The lower state, however, can be better populated by heating the cell, perhaps to 100°C, and using longer cells to enhance the absorption.

With extensive possibilities for frequency-stabilizing lasers to various absorbers, there is an international agreement regarding frequency values and uncertainties adopted for such lasers in respect of their application to the practical realization of the metre. This list is the *mise en pratique* of the Comité International des Poids et Mesures or CIPM [37] and includes a number of iodine-stabilized lasers, a laser diode based on the Rb two-photon absorption at 778 nm, an acetylene-stabilized diode at 1500 nm, together with some frequency references based on cooled atoms and ions. In this chapter, we will limit discussions to vapour-or gas-cell-based standards based on iodine, the 778-nm rubidium absorption and 1.5-μm acetylene absorptions.

The earliest iodine-stabilized He-Ne laser was based on the 11–5 R(127) absorption line in $^{127}I_2$, coincident with the He-Ne laser line [39]. This has been extensively characterized, so that various frequency shifts connected with iodine cell pressure, wall temperature, depth of FM and cavity geometry are well documented. The *mise en pratique* contains various parameter ranges within which the laser should be operated in order to obtain the specified reproducibility. In general, for a well-designed laser, around one part in 10^{11} reproducibility can be expected, although this is critically dependent upon the iodine cell quality. Systems can be made relatively 'user-friendly' through automatic in-lock monitoring,[3] because the $2f$ signal, which is at a maximum at line centre, can be used to indicate whether the laser is in lock. The $1f$ signal, which varies with the Doppler-limited background, can be used to indicate which of the main iodine components (d, e, f or g) the laser is locked to. An overall schematic diagram of a full system is shown in Figure 7.4. Third harmonic 633-nm iodine features are shown in Figure 7.9. There have been frequency chain measurements of this laser [40–42], as well as more recent measurements based on femtosecond laser combs [43,44]. Also, there have been direct international inter-comparisons between a number of countries on various different occasions (e.g. Ref. [45,46]). He-Ne lasers at 543, 612 and 640 nm can also be stabilized to Doppler-free iodine spectra, generally using iodine cells external to the cavity.

Solid-state lasers and diode lasers have also been stabilized to iodine. By far the most impressive results in terms of frequency stability have been obtained with the iodine-stabilized frequency-doubled Nd:YAG laser at 532 nm [47]. It can tune to a number of strong iodine transitions with narrow (≈1 MHz) linewidths and so this system can provide excellent frequency stability. The laser-to-laser reproducibility is usually determined by the iodine cell quality; background gases in the cell can cause small pressure shifts. An experimental arrangement similar to Figure 7.3 can be used, with an iodine cell length of a few tens of centimetres. There have been international inter-comparisons of this laser system, together with a number of frequency measurements based on femtosecond laser combs [48–51].

Laser diodes, operating in the range 630–690 nm, can also be frequency-stabilized to iodine but with increasing

---

[3] For a example of an automatically monitored iodine stabilised laser system at 633 nm see, http://www.npl.co.uk/length/wss/services/equipment/i2henelaser.html.

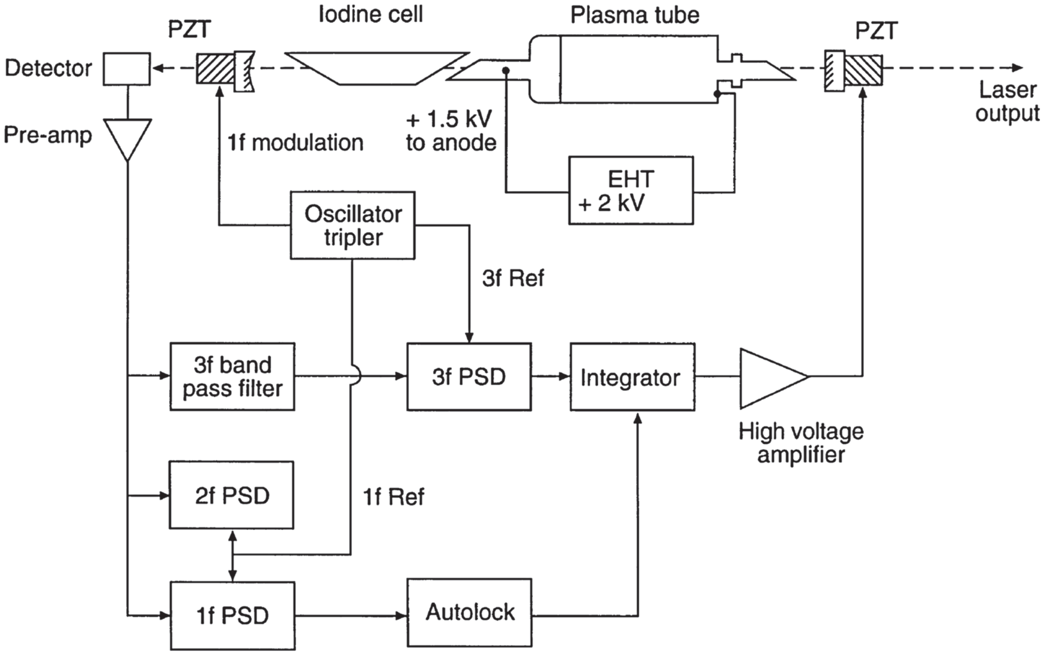


FIGURE 7.4 Overall schematic diagram for stabilizing a He-Ne laser to iodine, including the 1*f* signal required for component monitoring and the 2*f* signal used to advise the operator that the laser is in lock. PSD, phase-sensitive detector; PZT, piezoelectric.

difficulty for longer wavelengths. In principle, the tunability of laser diodes means that any one of hundreds of absorption lines could be used for frequency control. However, most interest has centred on those close in frequency to the 11–5 R(127) line used to stabilize the 633-nm He-Ne laser. One line often used is the 6–3 P(33) line [52], which has a stronger linear absorption than 11–5 R(127). Work has also been published on the R(60)8–4, R(125)9–4 and P(54)8–4 transitions. A number of countries took part in an international inter-comparison of eight lasers at the Bureau International des Poids et Mesures (BIPM) [53]. System reproducibility of a few tens of kHz was recorded, with best stabilities at 1 s of better than one part in 10^{11}.

The optical frequency standard at 778 nm is based on a two-photon transition in rubidium, whose partial energy level structure is shown in Figure 7.5. This standard has importance since it is a convenient reference at twice the frequency of the telecommunications region at 1556 nm. Two-photon spectroscopy is a Doppler-free technique, which can be used for some absorptions which are forbidden as conventional one-photon transitions. In two-photon spectroscopy, the two photons are absorbed, one from each of two counter-propagating beams. This method was first demonstrated in sodium [54]. In rubidium, the two-photon transition rate is enhanced because there is an intermediate level (the $^2P_{3/2}$ state) which is nearly halfway between the two levels involved in the two-photon transition. This is close to resonance with the ‘virtual state’ (shown dotted in Figure 7.5) reached after the absorption of just one photon. Detection of the transition is achieved by observing 420 nm fluorescence using a photomultiplier, which arises from the decay from the $P_{3/2}$ to the ground state. One difficulty with two-photon absorptions is that the signal is weak, although it increases as the square of the laser power. In order to maximize the signal, the cell is therefore placed inside a resonant power build-up cavity (Figure 7.6). Two servo systems are required: one to lock the cavity to the laser and the second to lock the laser to the transition peak. The cell is also heated to around 100°C in order to increase the rubidium vapour density. The frequency was first measured using a frequency chain [55,56], although femtosecond laser measurements are also planned.

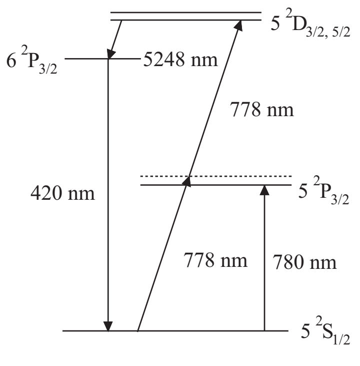


FIGURE 7.5 Partial-term scheme of rubidium, showing the two-photon transition at 778 nm, the decay route back to the ground state and the D-line transition at 780 nm.

The final system discussed here is a stabilized diode laser system based on saturated absorptions in acetylene [57]. A schematic diagram of a typical system is shown in Figure 7.7 and is also being developed at a number of standards laboratories (e.g. PTB, NPL and NRC). Acetylene has a number of molecular rotational and vibrational transitions between 1510 and 1540 nm for $^{12}C_2H_2$ and 1520–1550 nm for $^{13}C_2H_2$ and has importance through the provision of a number of standards covering the telecommunications window at 1500 nm.

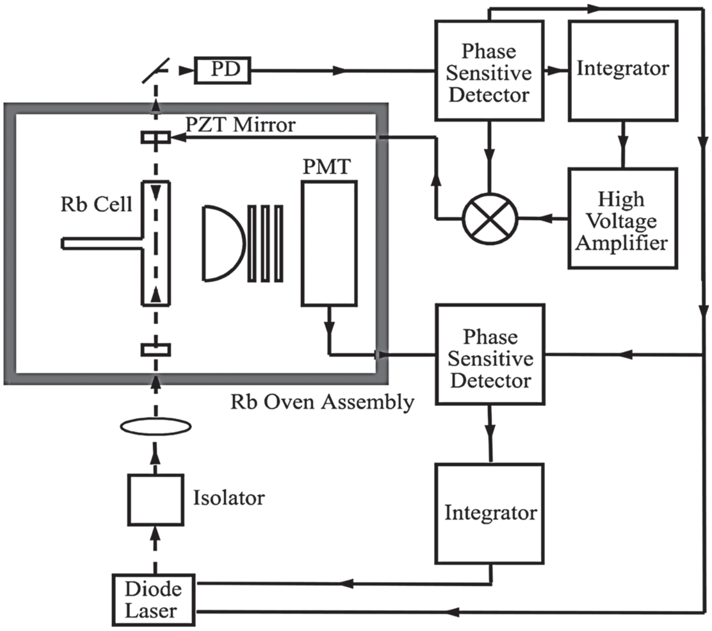

PD
Phase Sensitive Detector
Integrator
PZT Mirror
PMT
Rb Cell
High Voltage Amplifier
Phase Sensitive Detector
Rb Oven Assembly
Isolator
Integrator
Diode Laser

FIGURE 7.6 Schematic diagram of a system used to observe the two-photon transition at 778 nm. The photomultiplier tube (PMT) is used to record the 420 nm fluorescence [58].

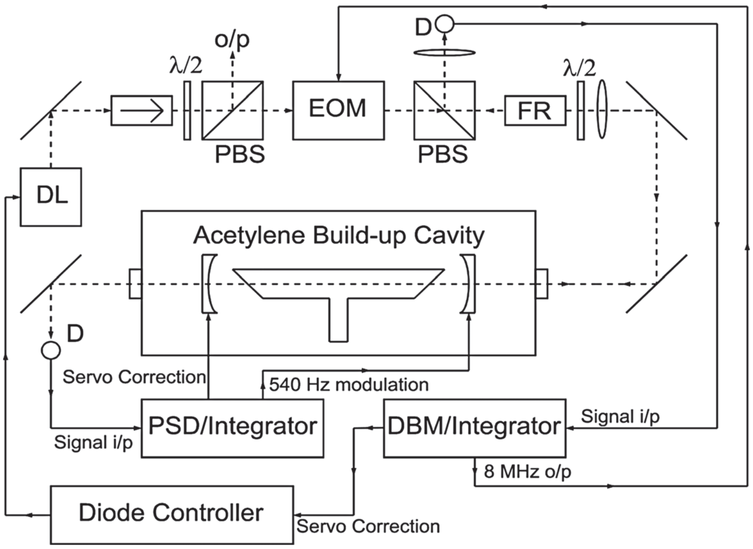

o/p
λ/2
EOM
D
λ/2
PBS
PBS
FR
DL
Acetylene Build-up Cavity
D
Servo Correction
540 Hz modulation
PSD/Integrator
DBM/Integrator
Signal i/p
Signal i/p
8 MHz o/p
Diode Controller
Servo Correction

FIGURE 7.7 Overall schematic diagram of the experimental arrangement used to observe Doppler-free transitions in acetylene Key: DL, diode laser, D, detector; EOM, electro-optic modulator; FR, Faraday rotator; PBS, polarizing beamsplitter, i/p, input; o/p, output and λ/2 denotes a half-wave plate [59].

With the ever-increasing demand for data transfer using fibre-based telecommunications, it is anticipated that fibres will eventually utilize the whole of the 1300–1600 nm region. Other molecular absorbers, such as CO, CH_4 or HCN, are able to provide frequency references in various other parts of this spectral region. Frequency standards based on some of these absorbers are being developed, with reproducibilities at a few parts in 10^{10} to one in 10^9 level for Doppler-free stabilization techniques. The acetylene system is a further variation of saturated absorption spectroscopy. In this implementation, the cell is inside a separate cavity to the laser. Again, the cavity transmission demonstrates an inverted Lamb dip at the centre of the Doppler absorption profile. In Figure 7.7, the laser is locked to the cavity and the cavity is then locked to the Doppler-free acetylene signal, using a third harmonic lock to remove the Doppler-limited background.

In acetylene, there is no hyperfine structure and so we only observe one peak within each Doppler-broadened transition. The $\nu_1 + \nu_3$ combination band of acetylene has more than 50 strong absorption lines in the 1510–1540 nm region.

All the frequency-stabilized laser systems discussed here have reproducibility limitations at the one part in 10^{11} level. There are a number of factors contributing to this limit and these are briefly discussed. The first is connected with the

Doppler shift. All the techniques discussed reduce or eliminate the first-order Doppler shift but are unable to reduce the second-order term, which arises from special relativity and varies as $v^2/2c^2$. In one dimension (for example, for motion along the laser beam), this relative frequency shift becomes $kT/2mc^2$, where k is the Boltzmann constant, T is the absolute temperature, m is the mass of the atom and c is the speed of light. Typically, this amounts to a few parts in 10^{12} of the optical frequency at room temperature, although this depends upon the mass of the absorbing atom or molecule. There will also be residual first-order Doppler shifts arising, for example, from the probe and saturating beams not being exactly co-linear. A further limitation will be frequency shifts caused by collisions, either with other molecules of the same type or with foreign (background) gases, produced in the cell preparation and sealing-off processes. These effects are difficult to quantify to better than the one part in 10^{11} level. There are also frequency broadening and shifts which can arise since the molecule only spends a finite time in the laser beam. These 'transit time' effects arise from the uncertainty principle: if the molecule only spends a time Δt in the beam, then the observed linewidth must be limited at the $1/\Delta t$ level in frequency. Finally, there are frequency-stability limitations arising from the signal-to-noise ratio of the saturated features. For a typical iodine linewidth of 5 MHz, a signal-to-noise ratio of around a thousand is required to achieve a stability of one part in 10^{11} (5 kHz). These limitations point to a requirement to have the frequency standard based on a transition in an atom which is cold (to remove Doppler effects and to keep the atom in the laser beam indefinitely), to be in a vacuum on its own (to remove collisional effects) and to have a transition with a low linewidth (to reduce the problems associated with the signal-to-noise ratio affecting the line centre). All these requirements point to the ideal of an ultimate frequency standard based on a cooled trapped atom or ion.

7.5 Evaluation of Frequency Stability and Reproducibility

The question arises on how to evaluate the performance of a stabilized laser system. The normal approach is to build two similar systems and then evaluate the difference (beat) frequency. This is typically arranged to be in the few tens of MHz to few GHz region, within the response range of commercially available detectors and amplifiers. The results are usually collected by a computer-controlled frequency counter; the computer is then also used for subsequent data analysis. A typical beat frequency monitoring system is shown in Figure 7.8.

The output beams from the two lasers should have the same (linear) polarization and approximately the same beam diameters and divergence in order to provide a strong rf signal. The two beams are overlapped at the beam combiner and focused onto the detector. The time-dependent part of the electric field of the laser beam may be expressed as $E\exp(\mathrm{i}\omega t)$, where the laser frequency is $\omega/2\pi$ and E is the electric field amplitude. The amplitude of the combined beams may, therefore, be represented by the expression $(E_1\exp(\mathrm{i}\omega_1 t) + E_2\exp(\mathrm{i}\omega_2 t))$. The detector measures the intensity of the combined beam:

$$\begin{aligned} I &= \left|E_1\exp(\mathrm{i}\,\omega_1 t) + E_2\exp(\mathrm{i}\,\omega_2 t)\right|^2 \\ &= E_1^2 + E_2^2 + 2E_1E_2\cos(\omega_1 + \omega_2)t \qquad (7.16) \end{aligned}$$

and so an rf beat frequency is produced at the difference frequency. The beat signal size depends not only on the laser intensities (as indicated earlier) but also upon the mode-matching, alignment and relative polarization of the two lasers. The beat is produced as a photocurrent in the detector which needs subsequent amplification.

Although one would expect the laser frequency noise to obey normal (Gaussian) statistics, with the standard deviation reducing as $1/\sqrt{(\text{measurement time})}$, this is not always the case. Often, the mean can vary over time as the laser drifts and a more appropriate statistic is a variance based on successive differences. In frequency metrology, this parameter is called the Allan variance, $\sigma^2(2, T, \tau)$ [60,61]. Here, τ is the averaging time (counter gate time) and T is the time between the start of successive readings. Ideally it is expected that $T=\tau$, but all frequency counters exhibit a 'dead time' between readings, especially if data need to be transferred to a computer. The statistic σ^2 is the variance of successive samples of the frequency averaged over time τ. If the output of the laser is represented by

$$E = E_0 exp\{j(\Omega t + \phi(t))\} \qquad (7.17)$$

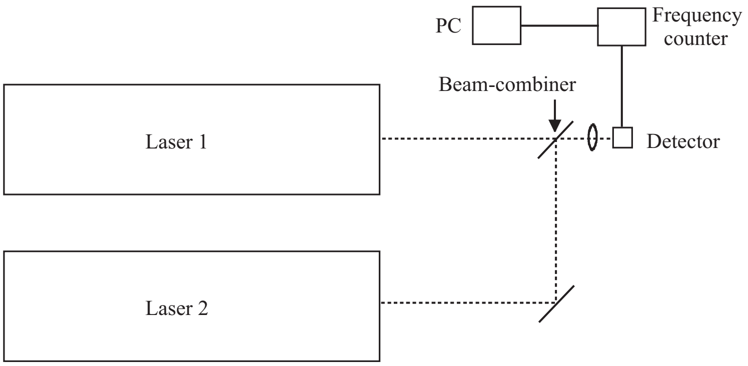


FIGURE 7.8 Measurement of the frequency difference between two lasers using rf beats.

then the instantaneous departure from the nominal frequency is $\dot{\phi}(t)/2\pi$. Hence, at time t, we have

$$\bar{f}_k = \frac{1}{2\pi\tau}\int_t^{t+\tau} \dot{\phi}(t)\,\mathrm{d}t. \tag{7.18}$$

If $\bar{f}_{k+1}$ and $\bar{f}_k$ are successive values of the frequency, averaged over time τ, then

$$\sigma^2(2,T,\tau) = \langle \frac{1}{2}(\bar{f}_{k+1} - \bar{f}_k)^2 \rangle \tag{7.19}$$

where <> denotes a time average. It may be shown that a log–log plot of σ *versus* τ (Allan deviation *versus* time) has various slopes, depending upon the type of frequency noise present. Investigation of the type of laser frequency noise present is one of the most important uses of this statistic. The most important σ *versus* τ slopes are listed in Table 7.1, together with the characteristic noise source yielding this slope.

Finally, the effect of counter 'dead time' should be noted. This can be very significant when $\tau \leq 30$ ms. However, during these periods, we are usually in a region where the slope is –1 (pure FM) or –1/2 (white frequency noise). In these cases, σ is unaffected by a non-zero counter dead time. A plot of a typical Allan deviation, in this case between two iodine-stabilized He-Ne lasers at 633 nm, is shown in Figure 7.9. The slope is approximately –1, typical for lasers which need to be frequency-modulated in order to produce the discriminants suitable for frequency stabilization.

TABLE 7.1

The Main Types of Laser Frequency Noise and Their Characteristic Allan Deviation Slopes

Slope	Source
–1	White phase noise or discrete FM
–1/2	White frequency noise
0	Flicker (1/*f*) walk
+1/2	Random frequency walk
+1	Steady frequency drift

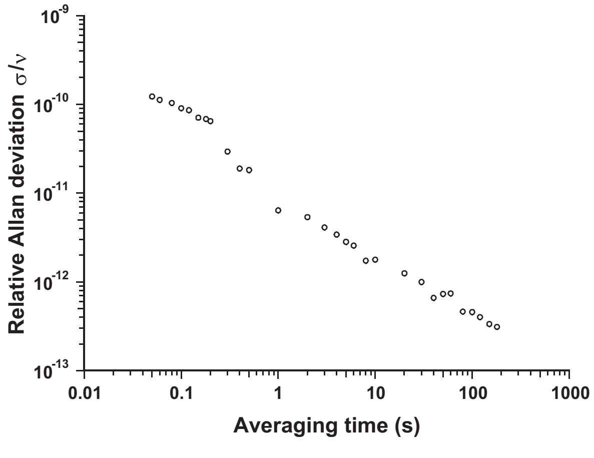


FIGURE 7.9 Plot of the Allan deviation of the beat frequency between two iodine-stabilized He-Ne lasers at 633 nm. The slope is –1, resulting from the FM of the lasers.

In addition to the measurement of frequency stability, we need to measure the mean offset between two laser systems to determine the long-term frequency reproducibility. If both lasers are emitting at nominally the same frequency, then the most convenient method is probably to shift one frequency by a known amount using an acousto-optic modulator and then compare the beat with the modulator drive frequency. However, if the lasers can be locked to different components in, for example, iodine that are a few tens of MHz apart, then there is an alternative method. Let us consider two components 'a' and 'b' and assume that $f_b > f_a$. We also assume that the components accessed by laser 1 are offset from the true value by Δf_1 and those accessed by laser 2 are offset by Δf_2. We lock laser 1 to 'a' and laser 2 to 'b', and the beat frequency (B) is then

$$B = (f_b + \Delta f_2) - (f_a + \Delta f_1). \tag{7.20}$$

If the components are exchanged, so that laser 2 is locked to 'a' and laser 1 to 'b', then the observed beat will become

$$B' = (f_b + \Delta f_1) - (f_a + \Delta f_2). \tag{7.21}$$

The unshifted interval is, therefore, $(B + B')/2$ and the difference between the laser offsets is given by $(B' - B)/2$. This difference between the offsets gives a measure of the long-term frequency reproducibility. This is usually a higher figure than the short-term frequency stability shown in Figure 7.9 and may result from small design differences between the two systems.

7.6 Cavity-Stabilization Techniques

Although atomic and molecular transitions provide the best long-term frequency stability and reproducibility for laser stabilization, an optical cavity can provide better short-term stability. This is because of the high signal-to-noise ratios obtainable with frequency discriminants based on optical cavity resonances observed either in reflection or transmission. Atomic transitions that are narrow are difficult to interrogate and lock onto if the laser linewidth is broad and the laser drifts through the transition too quickly. Typically, forbidden transitions in trapped ions, which are being developed as optical frequency standards, have linewidths of the order of 1 Hz and the observed linewidth is usually limited by the laser. A very high-finesse cavity that is acoustically well isolated and has a narrow fringe width and a low drift rate has a pivotal role in developing these standards. Lower finesse cavities provide an important means of stabilizing many different types of laser without the need for the lasing wavelength to be coincident with a quantum absorber. This is particularly relevant for tunable laser systems, such as dye lasers or Ti:sapphire solid-state lasers or many different diode lasers.

As with Doppler-free spectroscopy, there are a number of techniques available for frequency-locking to optical cavities. The simplest method is either to frequency-modulate the laser or to length-modulate the cavity at a few kHz and use a PSD to provide a first derivative feature in the same way as described earlier for locking to transitions in iodine. This error signal is then used, together with an integrator, to feed back onto the laser frequency. However, this method requires either the laser to be modulated or the cavity length to be dithered. Neither of these is desirable, since one often requires an unmodulated laser output, and similarly, modulating the cavity length can cause instabilities in the length caused by heating in the piezoelectrics used to effect tuning.

A method of frequency stabilization which does not require modulation is side-of-fringe locking [62]. With this technique, two signals are differenced in order to produce the error signal. The first is the transmitted fringe and the second is the power monitor. The result is a Fabry–Pérot fringe signal that is offset so that zero volts correspond to a frequency halfway up the side of the fringe. This lock point is immune to small power changes in the laser. However, this method can only give a small 'capture range'—the frequency range that the laser has to be within in order to acquire lock. This range will be extremely narrow for the case of ultrahigh-finesse cavities. Ideally, we require a lock feature that has a sharp change in signal with frequency at the cavity resonance line centre (that depends upon the cavity finesse) but still has a large capture range in order to acquire lock easily.

Another method which does not involve FM of the laser was first demonstrated by Hänsch and Coulliard [63]. In this method, linearly polarized light is reflected by a confocal reference cavity with an internal Brewster plate (or another optical element to induce a loss for one polarization). The incoming light comprises light polarized both parallel and perpendicular to the axis of the Brewster plate. One polarization, therefore, sees a cavity of low loss and experiences a frequency-dependent phase shift on reflection. The other component, to first order, is merely reflected by the input mirror. At resonance, both reflected beams are in phase and the reflected beam remains linearly polarized. Away from resonance, however, the parallel component acquires a phase shift and the reflected beam becomes elliptically polarized. This ellipticity can be detected with a simple polarization analyzer to give an error signal which has a dispersion shape, but with wings that extend the 'capture range' of the lock.

By far, the best cavity-lock method uses phase modulation of the laser to produce sidebands equation (7.15). Since the modulation is applied externally to the laser, we can pick off part of the beam before the modulator to provide a narrow-linewidth unmodulated source for high-resolution spectroscopy or frequency standards work. This method is normally referred to as the 'Pound–Drever–Hall' method [64] and produces a feature for locking similar to that obtained in FM spectroscopy. Close to cavity resonance, an imbalance is produced in the sidebands in both the reflected and transmitted powers and this produces amplitude modulation of the light at the detector. This signal is demodulated in a DBM to produce electronic signals suitable for locking. In order to achieve the narrowest linewidths, cavities with very high finesses are required. Typically, these cavities have a free spectral range of the order of 1 GHz and a finesse of 2×10^5. The mirrors must have very good surface quality, in order to reduce losses caused by scattering. If the mirrors cause significant scatter, then the cavity may have a high finesse but the transmission on resonance will be unacceptably low. The technology for producing these 'super-polished' mirrors was developed initially for aerospace applications and the production of laser gyros. Typical cavities have linewidths of a few kHz. This poses a problem in that the cavity has a characteristic response time, which is of the order of the reciprocal of the linewidth. If the laser frequency changes too quickly, then the transmitted power through the cavity will take time to change. The normal method of stabilization to these high-finesse cavities is, therefore, to use them in reflection. On time scales faster than the cavity response time, the reflected power is a combination of the light directly reflected from the cavity and the stored power inside the cavity. On these and faster time scales, the Pound–Drever–Hall lock, therefore, produces a phase error signal, rather than the frequency error signal relevant to longer time scales. A typical system is shown in Figure 7.10, and a typical lineshape observed from the DBM is shown in Figure 7.11.

There are two critical elements in producing a narrow-linewidth laser source. The first is to ensure that the electronic servocontrol scheme is capable of producing a narrow-linewidth laser. The usual method to determine the linewidth is to lock two lasers to adjacent modes of the same cavity and monitor the beat, which is, of course, at a frequency equivalent to one free spectral range. This usually produces very narrow linewidths, for example, 50 mHz has been reported [65], but there is a degree of common-mode noise rejection, so that the beat width is artificially narrowed. The more demanding task is to demonstrate narrow linewidths by locking two lasers to two independent cavities and measure the linewidth of the beat between them. It is necessary to acoustically isolate the cavity sufficiently in order to achieve this. A laser linewidth of only 0.2 Hz for averaging times up to 32 s, with a frequency

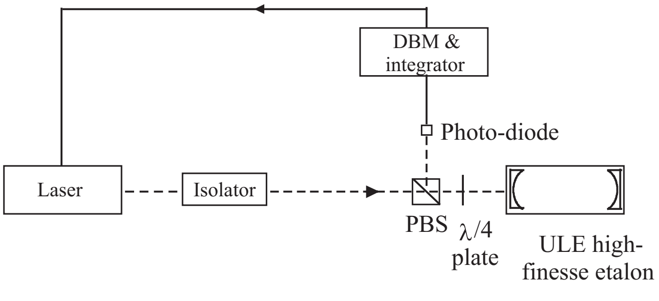


FIGURE 7.10 Typical arrangement for locking a laser to a cavity using the Pound–Drever–Hall method: PBS, polarizing beamsplitter; ULE, ultralow-expansion.

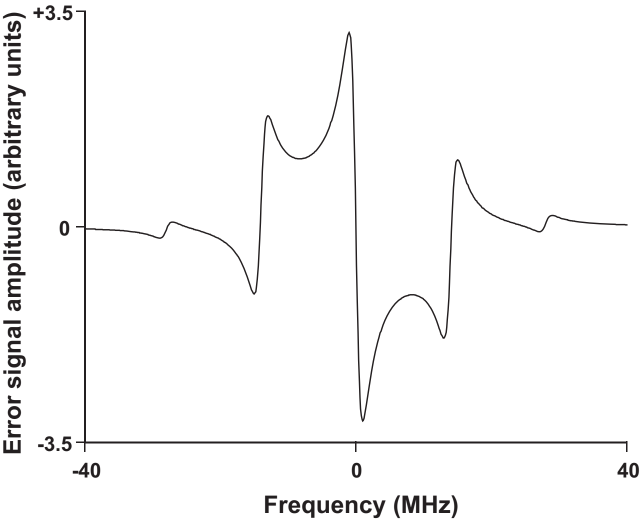


FIGURE 7.11 Calculated frequency discriminant, using a Pound–Drever–Hall lock using a 14 MHz modulation frequency and a cavity linewidth of 1 MHz and modulation index of 0.8. Note that this allows the laser to acquire lock over a frequency range determined by the modulation frequency (i.e. ±14 MHz); the central slope is determined by the cavity linewidth.

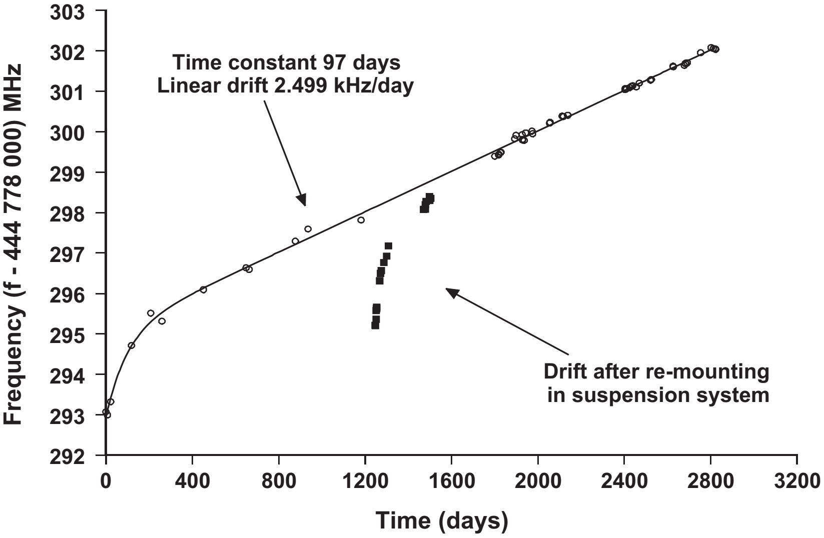


FIGURE 7.12 Long-term drift of an ULE cavity measured over several years.

stability of 3×10^{-16} at 1 s for a laser at 563 nm, was developed at NIST in order to interrogate their mercury trapped-ion standard [66, 67].

Although acoustic isolation is important to obtain short-term frequency stability, thermal stability and isothermal mechanical stability determine the longer-term drift. Typically, the ultralow-expansion (ULE) material used in their cavities has a thermal expansion coefficient of around $1 \times 10^{-8}\,^{\circ}\mathrm{C}^{-1}$. This coefficient is temperature-dependent and crosses through zero, at some temperature close to room temperature dependent upon the batch of material. If the cavity can be controlled close to this zero thermal expansion temperature, then this is clearly advantageous. In addition to thermally induced drift, longer-term isothermal drifts of cavities have been reported [68]. Over several years, a cavity isothermal drift rate of only 2.5 kHz/day ($\approx 0.03\,\mathrm{Hz\ s^{-1}}$) has been reported, as shown in Figure 7.12. These cavities thus provide excellent ‘optical flywheels’, providing good short-term frequency stability, whilst a quantum standard is used to provide the long-term frequency control.

7.7 Summary

This chapter has described a number of methods for frequency-stabilizing lasers. Relatively simple methods are available for 633-nm He-Ne lasers. Some systems are commercially available, which provide a reproducibility of one to a few parts in 10^8 of the optical frequency. For other lasers, modest frequency control is also possible using Doppler-limited absorption but the best frequency control for gas-cell-based reference lasers uses Doppler-free spectroscopy. If good short-term frequency stability is required but long-term reproducibility is less critical, then an ultra-stable cavity provides a convenient solution. Applications for these systems include dimensional metrology, gravitational wave detection, spectroscopy and the determination of fundamental constants.

REFERENCES

1. Kogelnik H. and Li T. 1966 *Proc. IEEE* **54** 1312–29.
2. Camparo J. C. 1985 *Contemp. Phys.* **26** 443–77.
3. Henry C. H. 1982 *IEEE J. Quantum Electron.* **QE-18** 259–64.
4. Welford D. and Mooradian A. 1982 *Appl. Phys. Lett.* **40** 865–7.
5. Johnston T. F. Jr, Brady R. H. and Profitt W. 1982 *Appl. Opt.* **21** 2307–16.
6. Gill P. 1993 Laser interferometry for precision engineering metrology *Optical Methods in Engineering Metrology* ed D. C. Williams (London: Chapman and Hall) ch 5, p. 153.
7. Birch K. P. and Downs M. J. 1994 *Metrologia* **31** 315–6.
8. Bönsch G. and Potulski E. 1998 *Metrologia* **35** 133–9.
9. Ciddor P. E. 1996 *Appl. Opt.* **35** 1566–73.
10. Birch K. P. and Downs M. J. 1993 *Metrologia* **30** 155–62.
11. Barwood G. P., Gill P. and Rowley W. R. C. 1993 *Meas. Sci. Technol.* **4** 988–94.
12. Barwood G. P., Gill P. and Rowley W. R. C. 1998 *Meas. Sci. Technol.* **9** 1036–41 and references therein.
13. Ju L., Blair D. G. and Zhao C. 2000 *Rep. Prog. Phys.* **63** 1317–427.
14. Drever R. W. P., Ford G. M., Hough J., Kerr I., Munley A. J., Pugh J. R., Robertson N. A. and Ward H. 1983 *Proc. 9th Int. Conf. on General Relativity and Gravitation* ed E. Shmutzer (Cambridge: Cambridge University Press).
15. Barwood G. P., Rowley W. R. C., Gill P., Flowers J. L. and Petley B. W. 1991 *Phys. Rev. A* **43** 4783–90.
16. Niering M. et al. 2000 *Phys. Rev. Lett.* **84** 5496–9.
17. Flowers J. L., Klein H. A., Knight D. J. E. and Margolis H. S. 2001 Hydrogenic systems for calculable frequency standards: status and options *NPL Report* CBTLM 11. This can be found at http://www.npl.co.uk/~jlf/refs/H_report_final_gs_processed.pdf.
18. Flowers J. L. and Petley B. W. 2001 *Rep. Prog. Phys.* **64** 1191–246.
19. Webb J. K. et al. 2001 *Phys. Rev. Lett.* **89** 091301.
20. Carilli C. L. et al. 2000 *Phys. Rev. Lett.* **85** 5511–14.
21. Prestage J. D., Tjoelker R. L. and Maleki L. 1995 *Phys. Rev. Lett.* **74** 3511.
22. Carilli C. L. 2001 *Phys. World.* **14** 26–7.
23. Balhorn R., Kunzmann H. and Lebowsky F. 1972 *Appl. Opt.* **11** 742.
24. Bennett S. J., Ward R. E. and Wilson D. C. 1973 *Appl. Opt.* **12** 1406.
25. Rowley W. R. C. and Gill P. 1990 *Appl. Phys. B* **51** 421–6.
26. Baer T., Kowalski F. V. and Hall J. L. 1980 *Appl. Opt.* **19** 3173–7.
27. Rowley W. R. C. 1990 *Meas. Sci. Technol.* **1** 348–51.
28. Brassington D. J. 1995 Tunable diode laser absorption spectroscopy for the measurement of atmospheric species *Spectroscopy in Environment Science (Advances in Spectroscopy 24)* ed R. J. H. Clark and R. E. Hester (Chichester: Wiley) ch 3.
29. Gilbert S. L., Swann W. C. and Dennis T. 2001 *Proc. SPIE* **4269** 184–91.
30. Swann W. C. and Gilbert S. L. 2000 *J. Opt. Soc. Am. B* **17** 1263–70.
31. Schulthess M. and Giggenbach D. 1998 *Electron. Lett.* **34** 1854.
32. Madej A. A., Marmet L. and Bernard J. E. 1998 *Appl. Phys. B* **67** 229–34.
33. Bjorklund G. C. 1980 *Opt. Lett.* **5** 15–17.
34. Drever R. W. P., Hall J. L., Kowalski F. V., Hough J., Ford G. M., Manley A. J. and Ward H. 1983 *Appl. Phys. B* **31** 97.
35. Hall J. L., Hollberg J. L., Baer T. and Robinson H. G. 1981 *Appl. Phys. Lett.* **39** 680.
36. Hall J. L., Robinson H. G., Baer T. and Hollberg L. 1983 The lineshapes of sub-Doppler resonances observable with FM sideband (optical heterodyne) laser techniques *Advances in Laser Spectroscopy (NATO Advanced Science Institutes Series)* ed F. T. Arrechi, F. Strumia and H. Walther (New York: Plenum).
37. Quinn T. J. 2003 *Metrologia* **40** 103–33.
38. Broyer M., Lehmann J-C. and Vigue J. 1975 *J. Physique* **36** 235–41.
39. Hanes G. R. and Dahlstrom C. E. 1969 *Appl. Phys. Lett.* **14** 362.
40. Acef O. et al. 1993 *Opt. Commun.* **97** 29–34.
41. Bernard J. E., Madej A. A., Siemsen K. J. and Marmet L. 2001 *Opt. Commun.* **187** 211–8.
42. Jennings D. A. et al. 1983 *Opt. Lett.* **8** 136–8.
43. Lea S. N. et al. 2003 *Metrologia* **40** 84–8.
44. Ye J. et al. 2000 *Phys. Rev. Lett.* **85** 3797–800.
45. Chartier J-M. and Chartier A. 2001 *Proc. SPIE* **4269** 123–33.
46. Darnedde H. et al. 1999 *Metrologia* **36** 199–206.
47. Ye J., Ma L-S. and Hall J. L. 2001 *Phys. Rev. Lett.* **87** 270801.
48. Robertsson L., Picard S., Hong F. L., Millerioux Y., Juncar P. and Ma L-S. 2001 *Metrologia* **38** 567–72.
49. Zhang Y., Ishikawa J. and Hong F-L. 2001 *Opt. Commun.* **200** 209–15.
50. Holzwarth R. et al. 2001 *Appl. Phys. B* **73** 269–71.
51. Rovera G. D., Ducos F., Zondy J-J., Acef O., Wallerand J-J., Knight J. C. and Russell P. St. J. 2002 *Meas. Sci. Technol.* **13** 918–22.
52. Edwards C. S., Barwood G. P., Gill P. and Rowley W. R. C. 1999 *Metrologia* **36** 41–5.
53. Zarka A. et al. 2000 *Metrologia* **37** 329–39.
54. Pritchard D., Apt J. and Ducas T. W. 1974 *Phys. Rev. Lett.* **32** 641–2.
55. Bernard J. E. et al. 2000 *Opt. Commun.* **173** 357–64.
56. Poulin M., Latrasse C., Touahri D. and Tetu M. 2002 *Opt. Commun.* **207** 233–42.
57. Onae A. et al. 1999 *IEEE Trans. Instrum. Meas.* **48** 563–6.
58. Edwards C. S., Barwood G. P., Gill P. and Rowley W. R. C. 2002 *Proc. 6th Symposium on Frequency Standards and Metrology* ed P. Gill (River Edge, NJ: World Scientific) pp. 643–5.
59. Edwards C. S., Barwood G. P., Gill P. and Rowley W. R. C. 2002 *Technical Digest of the LEOS 2002 Conference and Annual Meeting (Glasgow)* vol 1, pp. 281–2 (IEEE cat. no. 02CH37369)
60. Allan D. W 1966 *Proc. IEEE* **54** 221–30.

61. Allan D. W. 1987 *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **UFFC-34** 647–54.
62. Barger R. L., Sorem M. S. and Hall J. L. 1973 *Appl. Phys. Lett.* **22** 573.
63. Hänsch T. W. and Coulliard B. 1980 *Opt. Commun.* **35** 441.
64. Hall J. L., Baer T., Hollberg L. and Robinson H. G. 1981 *Laser Spectroscopy (Springer Series in Optical Sciences 30)* vol V, ed A. R. W. McKellar, T. Oka and B. P. Stoicheff (Berlin: Springer) p. 15.
65. Hils D. and Hall J. L. 1988 *Proc. 4th Symposium on Frequency Standards and Metrology (Ancona, Italy)* (New Jersey: World Scientific) pp. 162–73.
66. Young B. C., Cruz F. C., Itano W. M. and Bergquist J. C. 1999 *Phys. Rev. Lett.* **82** 3799–802.
67. Young B. C., Rafac R. J., Beall J. A., Cruz F. C., Itano W. M., Wineland D. J., and Bergquist J. C. 1999 *Proc. XIV Int. Conf. on Laser Spectroscopy* ed R. Blatt et al. (Singapore: World Scientific) pp. 61–70.
68. Barwood G. P., Gao K., Gill P., Huang G. and Klein H. A. 2001 *Proc. SPIE* **4269** 134–42.

FURTHER READING

Brassington D. J. 1995 Tunable diode laser absorption spectroscopy for the measurement of atmospheric species *Spectroscopy in Environment Science (Advances in Spectroscopy 24)* ed R. J. H. Clark and R. E. Hester (Chichester: Wiley) ch 3.

A review of linear absorption spectroscopy, from an environmental monitoring viewpoint.

Camparo J. C. 1985 *Contemp. Phys.* **26** 443–77.

A useful early review article, describing the general properties of laser diodes, for spectroscopic applications

Flowers J. L. and Petley B. W. 2001 *Rep. Prog. Phys.* **64** 1191–246.

A comprehensive and the review of fundamental constants (not only those such as the Rydberg and fine structure constants, which can be measured spectroscopically).

Gill P. 1993 Laser interferometry for precision engineering metrology *Optical Methods in Engineering Metrology* ed D. C. Williams (London: Chapman and Hall) ch 5, p. 153.

A review of the use of laser interferometry in dimensional metrology.

8

Laser Cooling and Trapping

Charles Adams and Ifan Hughes

CONTENTS

8.1 Introduction 127
8.2 Theory of Atom–Light Interactions 128
    8.2.1 Incoherent Excitation: Einstein A and B Coefficients 128
    8.2.2 Momentum Associated with Atom–Light Interactions: Radiation Pressure 128
    8.2.3 Coherent Excitation 129
        8.2.3.1 The Light Shift 129
    8.2.4 The Rabi Solution 130
    8.2.5 The Optical Bloch Equations 130
8.3 Light Forces 131
    8.3.1 The Dipole Force 132
    8.3.2 The Spontaneous Force 132
    8.3.3 Deceleration of an Atomic Beam 132
8.4 Doppler Cooling 132
    8.4.1 One-Dimensional Doppler Cooling—Two Counter-Propagating Beams 132
    8.4.2 Equilibrium Temperature 133
    8.4.3 Optical Molasses in Three Dimensions 134
8.5 Sub-Doppler Cooling 134
8.6 Trapping of Cold Atoms 135
    8.6.1 The Magneto-Optical Trap 135
    8.6.2 Magnetic and Optical Dipole Traps 136
8.7 Laser-Cooling Technology 136
8.8 Summary 137
References 137
Further Reading 138

8.1 Introduction

Of the many applications of lasers, the one with possibly the most dramatic impact on fundamental physics is laser cooling and trapping: the Nobel Prize for physics was awarded for laser cooling in 1997 [1]. The staggering progress achieved by laser cooling is reflected by considering a thought experiment that begins 'take an atom'. Such an experiment was impossible to realize in practice because room temperature atoms move at supersonic speeds, and as atoms are neutral, there was no easy way to grab hold of them. Laser cooling allows one to obtain samples of atoms essentially at rest, with temperatures in the micro-Kelvin region being routine. At such low temperatures, it becomes easy to trap or move atoms with modest electromagnetic forces [2,3] and a whole new class of experiments previously confined to the realm of fantasy becomes possible. For example, one can throw atoms upwards and use precision spectroscopy to measure time or gravity with unprecedented accuracy. One can place atoms at a desired location on a surface creating a novel lithographic technique or place an atom just above the surface to create a qubit in a quantum information processor. One can take many atoms and cool them further by evaporation to form a Bose–Einstein condensate [4] or degenerate Fermi gas. This last development has propelled atomic and optical physics right into the heart of condensed matter physics and resulted in the award of the Nobel Prize in 2001.

In this chapter we discuss the central ideas behind laser cooling and trapping of neutral atoms.[1] In Section 8.2 we introduce the theory of two-level atoms interacting with a near-resonant electromagnetic field. In Section 8.3 we discuss light forces. In Section 8.4 we explain Doppler cooling. In Section 8.5 we briefly describe the underlying principle of sub-Doppler cooling. In Section 8.6 we explain the principle of the magneto-optical trap. Finally, in Section 8.7 we describe some of the laser technologies required for laser-cooling experiments. An overview of other topics and applications of laser cooling can be found in the general references [8,9].

---

[1] The technology associated with trapping ions is rather different because ions have a strong interaction with electric fields allowing confinement without any pre-cooling (see Ref. [5,6]). However, laser cooling of trapped ions and tightly confined atoms, for example, in an optical lattice [7] are similar.

8.2 Theory of Atom–Light Interactions

First, we consider the case of a two-level system (atom, molecule) interacting with a resonant monochromatic light field. After reviewing the Einstein treatment of the two-level atom (incoherent excitation), we will consider coherent excitation.

8.2.1 Incoherent Excitation: Einstein A and B Coefficients

In his paper published in 1917 Einstein established the basic principles of atom–light interactions [10]. He postulated that radiative processes obeyed a set of rate equations relating the rate of change of state populations to their instantaneous values. He distinguished three basic processes: spontaneous emission, absorption and stimulated emission. The first two were familiar at the time, Einstein introduced the third. By considering the rate equations for atoms in equilibrium at a temperature T, Einstein determined the relationship between the coefficients corresponding to spontaneous and stimulated processes, A and B, respectively. If the ground and excited states have degeneracies g_a and g_b and energies E_a and E_b, respectively, then

$$g_a B_{ab} = g_b B_{ba} \text{ and } A = \frac{\hbar\omega_{ba}^3}{\pi^2 c^3} B_{ba} \tag{8.1}$$

with $\hbar\omega_{ba} = E_b - E_a$. As the coefficients are related, only one parameter is needed to characterize the strength of the atom–light interaction. This is also true in the quantum theory (which has the advantage that the Einstein coefficients can be derived). If the level degeneracies are equal, $g_b = g_a$, the fraction of atoms in the excited state in equilibrium is $N_b/(N_b + N_a) = (B_\rho)/(A + 2B_\rho)$, where ρ is the spectral energy density on resonance. In the limit of high intensity, this simplifies to $N_b/(N_b + N_a) = 1/2$, i.e. half the atoms are in the excited state. No matter how hard we 'pump', we cannot create a population inversion—this is referred to as *saturation*. To extend this treatment to more levels, one writes a rate equation for each level including all spontaneous routes into and out of that state plus all stimulated terms. One can also include other processes such as collisions.

8.2.2 Momentum Associated with Atom–Light Interactions: Radiation Pressure

In his 1917 paper [10], Einstein also discussed the importance of conservation of momentum in atom–light interactions. At that time there was a school of thought that energy and momentum conservations were statistical laws and did not have to be valid for elementary binary encounters. It was Einstein who stressed that when a photon and an atom interact, energy and linear momentum (and angular momentum) must all be conserved. The link between energy of an electromagnetic wave and momentum was known to Maxwell and had been demonstrated in the Crookes' radiometer experiments of Lebedev, Nichols and Hull in 1901 (see Table 8.1). However, for light–atom interactions, the assumption was that atoms radiated spherical waves, and from symmetry, no net momentum was exchanged with the field. Einstein pointed out that atoms must recoil and change their velocity. This idea survives in a modern quantum mechanical perspective and the force due to photon recoil is referred to as radiation pressure.

The first experimental verification of Einstein's photon recoil concept was Compton scattering (1927 Nobel Prize), where a photon interacts with a loosely bound atomic electron. As the electron recoils, the scattered photon has a different energy manifested as a change in wavelength. The measured change was found to be consistent with a collision involving a particle with initial momentum $\hbar k$. The mechanical effect of light on atoms was first demonstrated experimentally by Frisch in 1933, who deflected a sodium atomic beam with a sodium lamp. However, the effect was small, because the source could not provide sufficient intensity of resonant light. The advent of the laser created intense monochromatic tunable light beams and opened up enormous potential for laser manipulation of atoms, molecules and small macroscopic objects. In fact, laser trapping of microscopic objects, known as optical tweezing, was the first to be exploited and has led to many important discoveries and applications, especially in biology [11].

Early experiments on laser manipulation of atoms concentrated on slowing an atomic beam. A travelling electromagnetic plane wave with wave vector $\boldsymbol{k}$ consists of photons with a momentum $\hbar\boldsymbol{k}$. During absorption a momentum $\hbar\boldsymbol{k}$ is transferred to the atom, i.e. the atom (of mass M) changes speed by an amount $\hbar k/M$ in a direction parallel to the light beam. If the excited atom decays by stimulated emission, the emitted photon carries away a momentum $\hbar k$ and the atom loses the

TABLE 8.1

Year	Description
1620	Kepler incorrectly introduces radiation pressure concept to account for the direction of comet tails
1670	Newton *versus* Huygens: corpuscular *versus* wave theory of light
1873	Maxwell, $p = E/c$
1875	Crookes' radiometer. Wrong direction![a]
1901	Lebedev, Nichols and Hull, radiometer right direction!
1916	Einstein: concept of photon recoil
1924	Compton scattering
1933	Frisch deflects Na beam with a discharge lamp
1960	Invention of the laser
1973	Ashkin traps a glass bead with a laser
1982	Phillips slows an atomic beam with a laser
1985	Chu laser cools atoms to <1 mK

[a] Crookes' radiometer (or light-mill) consists of four vanes each with a dark absorbing surface on one side and a light reflecting surface on the other. The vanes are mounted on a pivot in an evacuated glass bulb, such that under illumination, the differential light pressure between the two sides induces a rotation. However, in a poor vacuum, differential effusion from the two surfaces causes a rotation in the opposite direction to that expected from the light pressure alone.

velocity it had gained. However, for an atom undergoing spontaneous emission, the photon can be emitted into a wide range of angles with a distribution that is symmetric with respect to the laser propagation direction. Consequently, the net recoil from many spontaneous emission processes averages to zero. Thus the net effect of repeated absorption–spontaneous emission cycles is to gradually reduce the atom's speed in multiples of the recoil speed $\hbar k/M$. For sodium atoms illuminated on the D transition, the recoil speed is ~3 cm s^{-1}; therefore, a sodium atom at room temperature with initial speed ~300 m s^{-1} can be brought to rest by absorbing ~10^4 photons, which only takes ~1 ms!

8.2.3 Coherent Excitation

The high brightness of masers and lasers makes it possible to access a regime of atomic excitation for which Einstein's rate equations are inadequate. A complete description is given by solving the time-dependent Schrödinger equation. If the electromagnetic field is nearly resonant with only two levels in the atom, a two-state model provides an accurate prediction of the dynamics. In the electric dipole approximation the light–atom interaction is described by the dipole operator,

$$\hat{D} = -\mathbf{d} \cdot \varepsilon = e\varepsilon \cdot \mathbf{r} \tag{8.2}$$

where $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_0 \cos \omega t$ is the electric field of the laser. The two-state atom is described by a wave function

$$\psi(\mathbf{r}) = C_\mathrm{a}(t)\varphi_\mathrm{a}(\mathbf{r})\mathrm{e}^{-iE_\mathrm{a}t/\hbar} + C_\mathrm{b}(t)\varphi_\mathrm{b}(r)\mathrm{e}^{-iE_\mathrm{b}t/\hbar} \tag{8.3}$$

given by the solution to Schrödinger's equation

$$\mathrm{i}\hbar\partial_t\psi = \hat{H}_\mathrm{at}\psi + \hat{D}\psi \tag{8.4}$$

where H_at is the atomic Hamiltonian. Substituting (8.3) in (8.4), and using orthogonality to eliminate the spatial dependence, one finds

$$\begin{aligned} \mathrm{i}\dot{C}_\mathrm{a} &= \omega_\mathrm{R}\mathrm{e}^{\mathrm{i}\phi}\cos\omega t\mathrm{e}^{-\mathrm{i}\omega_\mathrm{ba}t}C_\mathrm{b} \\ \mathrm{i}\dot{C}_\mathrm{b} &= \omega_\mathrm{R}\mathrm{e}^{-\mathrm{i}\phi}\cos\omega t\mathrm{e}^{\mathrm{i}\omega_\mathrm{ba}t}C_\mathrm{a} \end{aligned} \tag{8.5}$$

where the Rabi frequency ω_R is defined as

$$\hbar\omega_\mathrm{R}\mathrm{e}^{\mathrm{i}\phi} = e\int\varphi_\mathrm{a}^*\varepsilon_0 \cdot \mathbf{r}\varphi_\mathrm{b}\mathrm{d}^3\mathbf{r} \tag{8.6}$$

where ϕ is a phase. In the electric dipole approximation where the field does not vary significantly over the spatial extent of the atom, we can rewrite this definition as

$$\hbar\omega_\mathrm{R}\mathrm{e}^{\mathrm{i}\phi} = -d_\mathrm{ab}\varepsilon_0 \cdot \hat{\mathbf{e}} \tag{8.7}$$

where $d_\mathrm{ab} = -e\int\varphi_\mathrm{a}^* r\varphi_b \mathrm{d}^3 r$ is the dipole matrix element and $\hat{\mathbf{e}}$ is a unit polarization vector for the light field. If $\omega \sim \omega_\mathrm{ba}$, we can make the rotating-wave approximation and neglect terms in $\omega + \omega_\mathrm{ba}$ giving

$$\mathrm{i}\dot{C}_\mathrm{a} = \frac{1}{2}\omega_\mathrm{R}\mathrm{e}^{\mathrm{i}\phi}\mathrm{e}^{\mathrm{i}\Delta t}C_\mathrm{b} \text{ and } \mathrm{i}\dot{C}_\mathrm{b} = \frac{1}{2}\omega_\mathrm{R}\mathrm{e}^{-\mathrm{i}\phi}\mathrm{e}^{-\mathrm{i}\Delta t}C_\mathrm{a} \tag{8.8}$$

where the detuning $\Delta = \omega - \omega_\mathrm{ba}$. Substituting $c_\mathrm{a} = \mathrm{e}^{-\mathrm{i}\Delta t/2}C_\mathrm{a}$ and $c_\mathrm{b} = {}^{\mathrm{ei}\Delta t/2}C_\mathrm{b}$ we have

$$\begin{aligned} i\dot{c}_a &= \frac{1}{2}\left(\Delta c_\mathrm{a} + \omega_\mathrm{R}\mathrm{e}^{\mathrm{i}\phi}c_\mathrm{b}\right) \\ i\dot{c}_\mathrm{b} &= \frac{1}{2}\left(\omega_\mathrm{R}\mathrm{e}^{-\mathrm{i}\phi}c_\mathrm{a} - \Delta c_\mathrm{b}\right). \end{aligned} \tag{8.9}$$

If the wave function ψ is written as a two-component vector,

$$\psi = \begin{pmatrix} c_\mathrm{a} \\ c_\mathrm{b} \end{pmatrix} \tag{8.10}$$

then (8.9) becomes $\mathrm{i}\partial_t\psi = H\psi$ with,

$$H = \frac{\hbar}{2}\begin{pmatrix} \Delta & \omega_\mathrm{R}\mathrm{e}^{\mathrm{i}\phi} \\ \omega_\mathrm{R}\mathrm{e}^{-\mathrm{i}\phi} & -\Delta \end{pmatrix} \tag{8.11}$$

where the diagonal and off-diagonal terms represent energy shifts and the coupling between states, respectively. Equation (8.11) is often referred to as the dressed-state Hamiltonian [12].

8.2.3.1 *The Light Shift*

The energy levels of an atom in an electromagnetic field are given by the eigenvalues of the atom–field Hamiltonian (8.11). For laser intensity, I, the eigen energies are

$$U_\pm(I) = \pm\frac{\hbar}{2}\sqrt{\Delta^2 + \omega_\mathrm{R}(I)^2} = \pm\frac{\hbar}{2}\Omega \tag{8.12}$$

where

$$\Omega = \sqrt{\Delta^2 + \omega_\mathrm{R}^2} \tag{8.13}$$

is known as the *effective Rabi frequency*: The Rabi frequency is related to intensity by

$$\omega_\mathrm{R}(I) = \Gamma\sqrt{I/2I_\mathrm{s}} \tag{8.14}$$

where for a two-state system, the saturation intensity, $I_\mathrm{s} = \pi hc/3\tau\lambda^3$, λ and τ are the wavelength of the transition and excited state lifetime, respectively, and $\Gamma = 1/\tau$. Substituting for the Rabi frequency, the shift of the energy level due to the field, i.e. the ac Stark shift or light shift, is given by

$$\Delta U_\pm = U_\pm(I) - U_\pm(0) = \pm\frac{\hbar}{2}\left[\Omega(I) - \Delta\right]. \tag{8.15}$$

For $|\Delta| > \omega_\mathrm{R}$,

$$\Delta U_\pm(I) = \pm\frac{\hbar\omega_\mathrm{R}^2}{4|\Delta|} = \pm\frac{\hbar}{4|\Delta|\tau^2}\frac{I}{2I_\mathrm{s}} \tag{8.16}$$

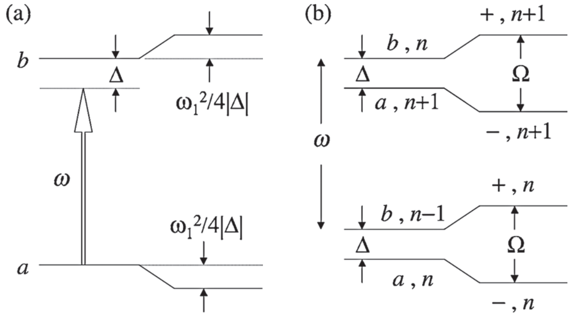


FIGURE 8.1 (a) The light shift for red detuning $\Delta < 0$. (b) The light shift in the dressed state picture [12].

The light shift is depicted schematically in Figure 8.1. An intensity gradient produces a potential gradient and, hence, a force. This is known as the *dipole force*, see Section 8.3.

8.2.4 The Rabi Solution

The solution to equation (8.9) is known as the Rabi solution. For the two-component vector $\psi(t) = [c_a(t), c_b(t)]$, the Rabi solution has the form

$$\psi(t) = \begin{pmatrix} \cos\Theta - i\dfrac{\Delta}{\Omega}\sin\Theta & -\mathrm{i}\dfrac{\omega_R}{\Omega}\sin\Theta \mathrm{e}^{\mathrm{i}\phi} \\ -\mathrm{i}\dfrac{\omega_R}{\Omega}\sin\Theta \mathrm{e}^{-\mathrm{i}\phi} & \cos\Theta + i\dfrac{\Delta}{\Omega}\sin\Theta \end{pmatrix} \psi(0) \tag{8.17}$$

where $\Theta = \Omega t/2$. Apart from the phase factor, the interaction is equivalent to a rotation of the state vector ψ in Hilbert space. In the absence of spontaneous emission, the theory of the two-state atom is identical to that of a magnetic spin in an oscillatory B-field (I. I. Rabi, Nobel Prize 1944).

For an atom initially in state a, $\psi(0) = (1, 0)$, excited on resonance,

$$\psi(t) = \begin{pmatrix} \cos\omega_R t/2 \\ -\mathrm{i}\mathrm{e}^{-\mathrm{i}\phi}\sin\omega_R t/2 \end{pmatrix} \tag{8.18}$$

i.e. the population oscillates at the Rabi frequency ω_R,

$$|c_b(t)|^2 = \sin^2(\omega_R t/2) \quad \text{and} \quad |c_a(t)|^2 = \cos^2(\omega_R t/2). \tag{8.19}$$

For $\Delta \neq 0$ the population oscillates at the effective Rabi frequency Ω.

8.2.5 The Optical Bloch Equations

The Bloch equations allow *dissipation*, i.e. spontaneous emission, to be included in the two-state model. They provide the foundations of nuclear magnetic resonance and magnetic resonance imaging, for which Bloch and Purcell were awarded the Nobel Prize in 1952. A wave function description of spontaneous emission is not feasible because information about the directions of emitted photons is lost. Instead one uses a density matrix. The density operator is defined as

$$\sigma = |\psi\rangle\langle\psi| \tag{8.20}$$

(as it involves a product of the wave function with its complex conjugate, the phase shifts which depend on the directions of spontaneous photons cancel). The density matrix is

$$\sigma = \psi\psi^* = \begin{pmatrix} C_a \\ C_b \end{pmatrix} \left(C_a^* C_b^* \right) = \begin{pmatrix} |C_a|^2 & C_a C_b^* \\ C_b C_a^* & |C_b|^2 \end{pmatrix}. \tag{8.21}$$

To abbreviate the notation, we define

$$\sigma = \begin{pmatrix} \sigma_{aa} & \sigma_{ab} \\ \sigma_{ba} & \sigma_{bb} \end{pmatrix} = \begin{pmatrix} |C_a|^2 & C_a C_b^* \\ C_b C_a^* & |C_b|^2 \end{pmatrix} \tag{8.22}$$

where σ_{aa} and σ_{bb} are the ground- and excited-state probabilities, respectively, and the off-diagonal terms σ_{ab} and σ_{ba} are known as coherences. As in optics, high coherence means that knowing the phase of one state allows us to determine that of the other. The rate of change of the density operator is given by the Liouville equation

$$\dot{\sigma} = \frac{\mathrm{i}}{\hbar}(\sigma H - H\sigma). \tag{8.23}$$

For convenience we introduce new variables,

$$\tilde{\sigma}_{bb} = \sigma_{bb} \quad \tilde{\sigma}_{ba} = \sigma_{ba}\mathrm{e}^{\mathrm{i}\omega t} \quad \tilde{\sigma}_{ab} = \sigma_{ab}\mathrm{e}^{-\mathrm{i}\omega t} \quad \tilde{\sigma}_{aa} = \sigma_{aa}. \tag{8.24}$$

Substituting into (8.23) we have:

$$\begin{aligned} \dot{\tilde{\sigma}}_{bb} &= -\mathrm{i}\frac{\omega_R}{2}(\tilde{\sigma}_{ab} - \tilde{\sigma}_{ba}) \\ \dot{\tilde{\sigma}}_{ba} &= \mathrm{i}\Delta\tilde{\sigma}_{ba} + \mathrm{i}\frac{\omega_R}{2}(\tilde{\sigma}_{bb} - \tilde{\sigma}_{aa}) \\ \dot{\tilde{\sigma}}_{ab} &= -\mathrm{i}\Delta\tilde{\sigma}_{ab} - \mathrm{i}\frac{\omega_R}{2}(\tilde{\sigma}_{bb} - \tilde{\sigma}_{aa}) \\ \dot{\tilde{\sigma}}_{aa} &= \mathrm{i}\frac{\omega_R}{2}(\tilde{\sigma}_{ab} - \tilde{\sigma}_{ba}). \end{aligned} \tag{8.25}$$

The effect of spontaneous emission is to reduce the excited-state population at a rate $\Gamma = 1/\tau$, and the coherences at a rate $\Gamma/2$. Adding these decay terms gives

$$\begin{aligned} \dot{\tilde{\sigma}}_{bb} &= -\mathrm{i}\frac{\omega_R}{2}(\tilde{\sigma}_{ab} - \tilde{\sigma}_{ba}) - \Gamma\tilde{\sigma}_{bb} \\ \dot{\tilde{\sigma}}_{ba} &= \mathrm{i}\Delta\tilde{\sigma}_{ba} + \mathrm{i}\frac{\omega_R}{2}(\tilde{\sigma}_{bb} - \tilde{\sigma}_{aa}) - \frac{\Gamma}{2}\tilde{\sigma}_{ba} \\ \dot{\tilde{\sigma}}_{ab} &= -\mathrm{i}\Delta\tilde{\sigma}_{ab} - \mathrm{i}\frac{\omega_R}{2}(\tilde{\sigma}_{bb} - \tilde{\sigma}_{aa}) - \frac{\Gamma}{2}\tilde{\sigma}_{ab} \\ \dot{\tilde{\sigma}}_{aa} &= \mathrm{i}\frac{\omega_R}{2}(\tilde{\sigma}_{ab} - \tilde{\sigma}_{ba}) + \Gamma\tilde{\sigma}_{bb}. \end{aligned} \tag{8.26}$$

The steady-state solution is found from $\sigma = 0$. For the population in the excited state, one finds

$$\sigma_{\mathrm{bb}} = \frac{\omega_{\mathrm{R}}^2/4}{\Delta^2 + \Gamma^2/4 + \omega_{\mathrm{R}}^2/2}. \tag{8.27}$$

This corresponds to a Lorentzian lineshape with width Γ in the limit of low intensity, $\omega_1 \rightarrow 0$. At higher intensity the lineshape becomes *power-broadened*. As the occupation probabilities must add up to one, $\sigma_{\mathrm{aa}} + \sigma_{\mathrm{bb}} = 1$, the number of independent parameters can be reduced to three. By defining

$$u = \frac{1}{2}\left(\tilde{\sigma}_{\mathrm{ab}} + \tilde{\sigma}_{\mathrm{ba}}\right) \quad v = \frac{1}{2\mathrm{i}}\left(\tilde{\sigma}_{\mathrm{ab}} - \tilde{\sigma}_{\mathrm{ba}}\right) \quad \omega = \frac{1}{2}\left(\tilde{\sigma}_{\mathrm{bb}} - \tilde{\sigma}_{\mathrm{aa}}\right) \tag{8.28}$$

Equation (8.26) becomes

$$\begin{aligned} \dot{u} &= -\frac{\Gamma}{2}u + \Delta v \\ \dot{v} &= -\Delta u - \frac{\Gamma}{2}v - \omega_{\mathrm{R}} w \\ \dot{w} &= \omega_{\mathrm{R}} v - \Gamma\left(w + \frac{1}{2}\right). \end{aligned} \tag{8.29}$$

These equations are known as *optical Bloch equations*. By definition $w + \psi$ is equal to σ_{bb}. The physical significance of u and v is illustrated by calculating the expectation value of the dipole moment

$$\begin{aligned} \langle \mathbf{d} \rangle &= \mathrm{Tr}(\sigma \mathbf{d}) = d_{\mathrm{ab}}\left(\tilde{\sigma}_{\mathrm{ab}} \mathrm{e}^{\mathrm{i}\omega t} + \tilde{\sigma}_{\mathrm{ab}} \mathrm{e}^{-\mathrm{i}\omega t}\right) \\ &= 2d_{\mathrm{ab}}\left(u \cos\omega t - v \sin\omega t\right). \end{aligned} \tag{8.30}$$

Comparison with $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_0 \cos \omega t$ suggests that u and v represent the in-phase and quadrature components of the average dipole moment. Neglecting the damping terms, equation (8.29) is equivalent to the precession of a Bloch vector $\rho = \{u, v, w\}$ under a torque $\mathbf{\Omega} = \{\omega_{\mathrm{R}}, 0, -\Delta\}$. The precession frequency is given by the length of the torque vector $\Omega = \sqrt{\omega_{\mathrm{R}}^2 + \Delta^2}$, which is equal to the *effective Rabi frequency*. The azimuthal and polar angles of the Bloch vector, $\phi = -\tan^{-1}(u/v)$ and $\theta = \tan^{-1}[(u^2 + v^2)^{1/2}/w)]$, characterize the phase of the light field and the population inversion, respectively. The steady-state solution of the optical Bloch equations is

$$u_{\mathrm{st}} = \frac{\Delta}{\omega_{\mathrm{R}}}\frac{s}{1+s} \quad v_{\mathrm{st}} = \frac{\Gamma}{2\omega_{\mathrm{R}}}\frac{s}{1+s} \quad w_{\mathrm{st}} = -\frac{1}{2}\frac{1}{1+s} \tag{8.31}$$

where $s = 2\omega_{\mathrm{R}}^2/\left(4\Delta^2 + \Gamma^2\right)$ is known as the saturation parameter. This result is used in Section 8.3 to derive an expression for the force exerted on an atom in a light field.

The main differences between the rate equations and the optical Bloch equations are the Rabi oscillations. Also worth noting is that the initial change in excitation probability varies linearly in time for rate-equation excitation and quadratically in time for coherent excitation. Consequently, even for short times, before oscillatory behaviour becomes apparent, coherent excitation yields a qualitatively different behaviour from the rate-equation approach. Two questions arise: can we regain the rate-equation results from the optical Bloch equations? When can we use rate equations? They are, in fact, valid and useful, when (i) the bandwidth of the incident light exceeds the atomic transition linewidth; or (ii) the combined collision and Doppler linewidth greatly exceed the radiative transition linewidth. In this case, one can use adiabatic elimination to solve for the fast time evolution of the coherences, leaving just the population changes. The range of validity of the rate equations matches well with experiments performed prior to the advent of the laser.

8.3 Light Forces

To obtain an expression for the mean force acting on a two-level atom in a near-resonant laser field we adopt a semi-classical approach [13]. The force on an atom due to a perturbation $\hat{D}$ is given by Ehrenfest's equation:

$$\mathbf{F} = \left\langle -\nabla \hat{D} \right\rangle = \left\langle \nabla(\mathbf{d} \cdot \boldsymbol{\varepsilon}) \right\rangle. \tag{8.32}$$

The gradient operator is taken with respect to $\boldsymbol{R}$, the position operator of the atomic centre of mass. In general, the atomic wavepacket is smaller than the wavelength of the light and we can treat $\boldsymbol{R}$ as a classical variable. As the atomic dipole does not depend on $\boldsymbol{R}$, we can write

$$\mathbf{F} = \langle \mathbf{d} \rangle \nabla \boldsymbol{\varepsilon} \tag{8.33}$$

We generalize the form of the electric field to include a position-dependent phase, χ, i.e.

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_0 \cos\left(\omega t + \chi\right). \tag{8.34}$$

Substituting for $\boldsymbol{\varepsilon}$ and the value of $\langle \boldsymbol{d} \rangle$ given by (8.30), the time-averaged force is

$$\langle \mathbf{F} \rangle = d_{\mathrm{ab}}(u_{\mathrm{st}} \nabla \varepsilon_0 + v_{\mathrm{st}} \varepsilon_0 \nabla \chi). \tag{8.35}$$

The force has two components: the first term

$$\mathbf{F}_{\mathrm{dip}} = d_{\mathrm{ab}}\left(u_{\mathrm{st}} \nabla \varepsilon_0 + \hbar u_{\mathrm{st}} \nabla \omega_{\mathrm{R}}\right). \tag{8.36}$$

is proportional to the gradient of the field and the real part of the atomic dipole, and is known as the *gradient force* or *dipole force*. The dipole force arises from the redistribution of photons in the light field by absorption and stimulated emission cycles. The second term

$$\mathbf{F}_{\mathrm{spont}} = d_{\mathrm{ab}} \varepsilon_0 v_{\mathrm{st}} \nabla \chi = -\hbar \omega_{\mathrm{R}} \nabla \chi \tag{8.37}$$

is proportional to the gradient of the phase and the imaginary part of the atomic dipole, and is known as the *radiation pressure force* or *spontaneous force* because it arises from absorption and spontaneous emission cycles.

8.3.1 The Dipole Force

Substituting for u_{st} from (8.31), the dipole force is

$$\boldsymbol{F}_{\mathrm{dip}} = -\frac{\hbar\Delta}{2}\frac{\omega_{\mathrm{R}}\nabla\omega_{\mathrm{R}}}{\Delta^2+\Gamma^2/4+\omega_{\mathrm{R}}^2/2}. \tag{8.38}$$

Using $\omega_{\mathrm{R}} = \Gamma\sqrt{I/2I_{\mathrm{s}}}$, this becomes

$$\boldsymbol{F}_{\mathrm{dip}} = -\frac{\hbar\Delta/2}{I+I_{\mathrm{s}}\left(1+4\Delta^2/\Gamma^2\right)}\nabla I \tag{8.39}$$

i.e. the dipole force is proportional to the gradient of the field intensity. The dipole force acts in the direction of the intensity gradient: towards low intensity ('weak-field seeking') for positive or blue detuning and high intensity ('strong-field seeking') for red detuning. For a red-detuned laser the induced dipole is in-phase with the driving field and an atom can lower its energy, $-\boldsymbol{d}\cdot\varepsilon = -d\varepsilon$, by moving to a region of high field. Whereas, for blue-detuning, the induced dipole moment oscillates out-of-phase with the field and the atom can lower its energy, $-\boldsymbol{d}\cdot\varepsilon = +d\varepsilon$, by moving to lower field. On resonance, the dipole is exactly $\pi/2$ out-of-phase with the field, and the average energy transfer per cycle is zero.

Further insight into the dipole force is provided by considering the dressed-state model [12]. In the limit of large detuning where spontaneous emission is negligible, the dipole force can be thought of as the gradient of dressed-state energies given by (8.16). In this limit the dipole force is conservative. We can also understand the force as arising from absorption-stimulated emission cycles. For example, in a standing wave an atom can absorb a photon from the left and emit to the right giving a net momentum transfer. Unlike the spontaneous force, the dipole force does not saturate. Large forces arise when the light intensity changes rapidly over a short distance, for example, in standing waves (optical lattices [7]), tightly focused laser beams and evanescent waves. Atom-optical components exploiting these geometries are used to manipulate cold atoms [2].

For large detuning the dipole force scales as $F \propto I/\Delta$, whereas for the spontaneous force, $F \propto I/\Delta^2$ (see later); therefore, in the limit of high intensity and large detuning, one can isolate the dipole force. This is particularly important for optical trapping of atoms where spontaneous emission can be an undesirable heating mechanism. A good example is far-off resonant optical dipole trapping of Bose–Einstein condensates.

8.3.2 The Spontaneous Force

For a plane wave $\varepsilon = \varepsilon_0\cos(\omega t - \boldsymbol{k}\cdot\boldsymbol{R})$, $\nabla\chi = -\boldsymbol{k}$ and the spontaneous force (8.37) is given by

$$\mathbf{F} = \hbar\mathbf{k}\omega_{\mathrm{R}}v_{\mathrm{st}} \tag{8.40}$$

Substituting for v_{st} from (8.31) and using $\omega_{\mathrm{R}} = \Gamma\sqrt{I/2I_{\mathrm{s}}}$, one finds

$$\mathbf{F} = \hbar\mathbf{k}\frac{\Gamma}{2}\frac{I}{2I+I_{\mathrm{s}}\left(1+4\Delta^2/\Gamma^2\right)}. \tag{8.41}$$

This result is simply the product of the photon scattering rate, $\Gamma\sigma_{\mathrm{bb}}$, and the net recoil per collision, $\hbar\boldsymbol{k}$. At low intensity, $\boldsymbol{F} \propto I$, while at high intensity, the force saturates $\boldsymbol{F} \to \hbar\boldsymbol{k}\Gamma/2$. Even though the photon momentum is small, the spontaneous force can result in a substantial acceleration ($10^5 g$), because photons may be scattered at a rate Γ (i.e. typically $10^8\,\mathrm{s}^{-1}$). The velocity dependence appears due to the *Doppler shift*. For an atom with velocity $\boldsymbol{v}$, the detuning becomes $\Delta = \omega - \omega_{\mathrm{ab}} - \boldsymbol{k}\cdot\boldsymbol{v}$, and the spontaneous force is

$$\mathbf{F}(\mathbf{v}) = \hbar\mathbf{k}\frac{\Gamma}{2}\left[\frac{I}{I+I_{\mathrm{s}}\left(1+4(\Delta-\mathbf{k}\cdot\mathbf{v})^2/\Gamma^2\right)}\right]. \tag{8.42}$$

The scattering force is dissipative and cannot be represented by a potential. The first important application of the spontaneous force was to slow an atomic beam [14].

8.3.3 Deceleration of an Atomic Beam

Consider a resonant laser beam that is counter-propagating with a co-linear atomic beam. An atom of mass M will slow down by a speed $\hbar k/M$ each time it absorbs a photon. For alkali metal atoms emerging from an oven heated to a few 100 degrees the stopping distance is about 1 m, and the stopping time is about a millisecond, so it is feasible to 'stop' a beam in a reasonable-sized vacuum chamber. However, as the atoms slow the Doppler shift changes and the spontaneous force is reduced. The capture range of the spontaneous force is $\Gamma/k \sim 5\,\mathrm{m\,s^{-1}}$ for alkali metal atoms, whereas the atomic beam velocity is a few hundred $\mathrm{m\ s^{-1}}$. Two tricks have been employed to keep the laser on resonance while the atom slows: (i) *chirp cooling* and (ii) *Zeeman slowing*. In chirped cooling the frequency of the laser is changed in time ('chirped') such that the acceleration remains close to a maximum for a pulse of atoms that 'catch' the chirp. In Zeeman slowing, the atoms enter a tapered magnet which Zeeman shifts the atomic levels such that the laser remains on resonance with the decelerating atoms. Note that repeated absorption–emission cycles are vital, and hence, it is crucial to illuminate atoms with light on a closed transition (see Section 8.7).

After the first successful slowing experiments in 1982 [14], it became the standard technique in obtaining cold atoms from a beam. However, although a single laser beam dramatically reduces $\langle v\rangle$ and can also reduce the temperature, $T \propto \langle v^2\rangle$, obtaining a nearly stationary cloud requires a light force component in each direction. This can be achieved by using three orthogonal pairs of counter-propagating laser beams, as we show in the next section.

8.4 Doppler Cooling

8.4.1 One-Dimensional Doppler Cooling—Two Counter-Propagating Beams

Consider an atom placed between counter-propagating laser beams, both with intensity, I. For $I < I_{\mathrm{s}}$, the total force is given by the sum due to each beam, i.e.

$$F = F_+ - F_- \tag{8.43}$$

where

$$F_\pm = \hbar k \frac{\Gamma}{2} \frac{I/I_s}{2I/I_s + 1 + 4(\Delta \mp kv)^2/\Gamma^2}. \tag{8.44}$$

The total force and that due to each component alone are plotted in Figure 8.2. Note that the intensity factor in the denominator is $2I$ because both beams contribute to saturation. For small velocities, $k|v| < \Gamma$, we may neglect terms in v^2 and expand the denominator up to first order in v giving

$$F = 4\hbar k^2 \frac{I}{I_s} \frac{2\Delta/\Gamma v}{\left(2I/I_s + 1 + 4\Delta^2/\Gamma^2\right)^2}. \tag{8.45}$$

Consequently for $\Delta < 0$ (red detuning), the atoms experience a *friction-like force*,

$$F = -\alpha v \tag{8.46}$$

where α is the friction coefficient. The rate of change of the kinetic energy is given by

$$\left(\frac{\mathrm{d}E}{\mathrm{d}t}\right)_{\text{cool}} = -\alpha v^2. \tag{8.47}$$

A positive friction coefficient leads to cooling. Due to the central role of the Doppler effect, this process is known as *Doppler cooling*. In three dimensions (three orthogonal counter-propagating beam pairs), the atoms are constantly slowed down towards $v = 0$, forming a treacle-like substance christened *optical molasses*.

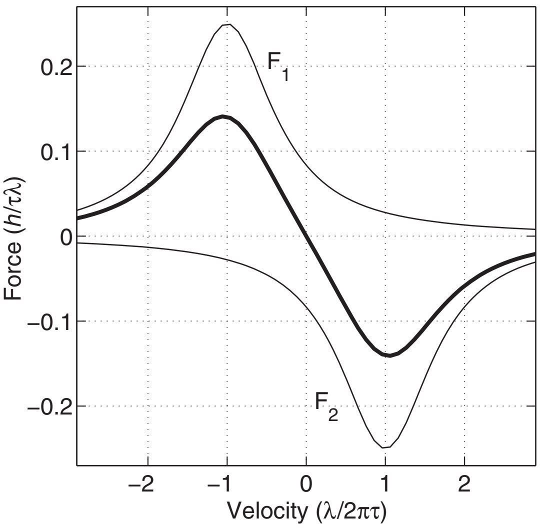


FIGURE 8.2 Velocity dependence of the spontaneous force for counter-propagating laser beams with $\Delta = -\Gamma$ and $I = I_s$. The force due to each beam alone is shown as a grey line. The effective capture range of the Doppler cooling is of the order of $\pm\Delta/k$.

8.4.2 Equilibrium Temperature

Superficially, it appears that atoms in optical molasses will be cooled to zero velocity and zero temperature. This does not happen because of a heating effect due to the random nature of spontaneous emission. The spontaneous recoils cause the atoms to undergo a random walk in momentum space. In one dimension, the mean square momentum increases as $\langle p^2 \rangle = 2D_p t$, where the momentum diffusion constant is equal to the square of the photon recoil times the scattering rate, i.e. $D_p = (\hbar k)^2 R$ [12]. The kinetic energy will increase at a rate

$$\left(\frac{\mathrm{d}E}{\mathrm{d}t}\right)_{\text{heat}} = \frac{(\hbar k)^2}{M} R. \tag{8.48}$$

In equilibrium the heating and cooling rates balance, thus, using the photon scattering rate

$$R = \frac{I/I_s}{\left(2I/I_s + 1 + 4\Delta^2/\Gamma^2\right)} \tag{8.49}$$

and the friction coefficient determined earlier,

$$\alpha = -4\hbar k^2 \frac{I}{I_s} \frac{2\Delta/\Gamma}{\left(2I/I_s + 1 + 4\Delta^2/\Gamma^2\right)^2} \tag{8.50}$$

we obtain

$$v^2 = \frac{\hbar\Gamma}{4M} \frac{1 + 2I/I_s\left(2\Delta/\Gamma\right)^2}{2\Delta/\Gamma}. \tag{8.51}$$

The expression for the heating and cooling rates is time or ensemble averages; therefore, we interpret this equation as the mean-square velocity of the ensemble or the time average for a single atom. Taking the thermal energy to be $k_B T$ per degree of freedom we obtain the following expression for the equilibrium temperature:

$$k_B T = -\frac{\hbar\Gamma}{4}\left[\frac{2\Delta}{\Gamma} + \frac{\Gamma}{2\Delta}\left(\frac{2I}{I_s} + 1\right)\right]. \tag{D4.7.52}$$

The minimum temperature is achieved with $I \ll I_s$ and a detuning of $\Delta = -\Gamma/2$, which yields

$$k_B T_D = \frac{\hbar\Gamma}{2}. \tag{8.53}$$

This is known as the Doppler-cooling limit. It is a very elegant result—the minimum temperature only depends on the atom's linewidth and (for low intensities) is independent of the intensity and detuning from resonance. The Doppler temperature is extremely low—of the order of hundreds of micro Kelvin for the alkali metal atoms! An important issue relates to the concept of temperature for laser-cooled atoms. A more elaborate treatment, using the Fokker–Planck equation, leads to the conclusion that the ensemble has a Maxwellian distribution of velocities, with a temperature given by (8.53). Equilibrium is attained (even without collisions between atoms) by the interaction with the light field.

8.4.3 Optical Molasses in Three Dimensions

Three-dimensional optical molasses using counter-propagating laser beams along three mutually orthogonal directions was first observed in 1985 by Chu et al. [15]. Care must be taken to avoid external magnetic fields or imbalances in the beam intensities as they lead to a net drift of atoms out of the cooling region. In contrast to Millikan's oil-drop experiment, the atoms in optical molasses are not held up by a force which just balances gravity. A constant force such as gravity induces a downward drift at a speed Mg/α. However, for typical parameters the magnitude of the viscosity is so high that the gravity-induced drift is only millimetres per second. As a result of light-force imbalance and stray fields, atoms can be held in the molasses region for time scales between tens of milliseconds and seconds. Longer storage times require the introduction of a confining potential, see Section 8.6.

By deliberately introducing a frequency difference between the laser beams, we can cool atoms in a moving frame (the frame where the Doppler shift compensates the frequency difference of the beams in the laboratory frame). This allows us to launch atoms with any mean velocity to form an atomic fountain. The fountain geometry provides the longest free-space interaction times and consequently is used in precision measurement experiments such as atomic clocks and gravitometry (see e.g. Ref. [8] and references therein).

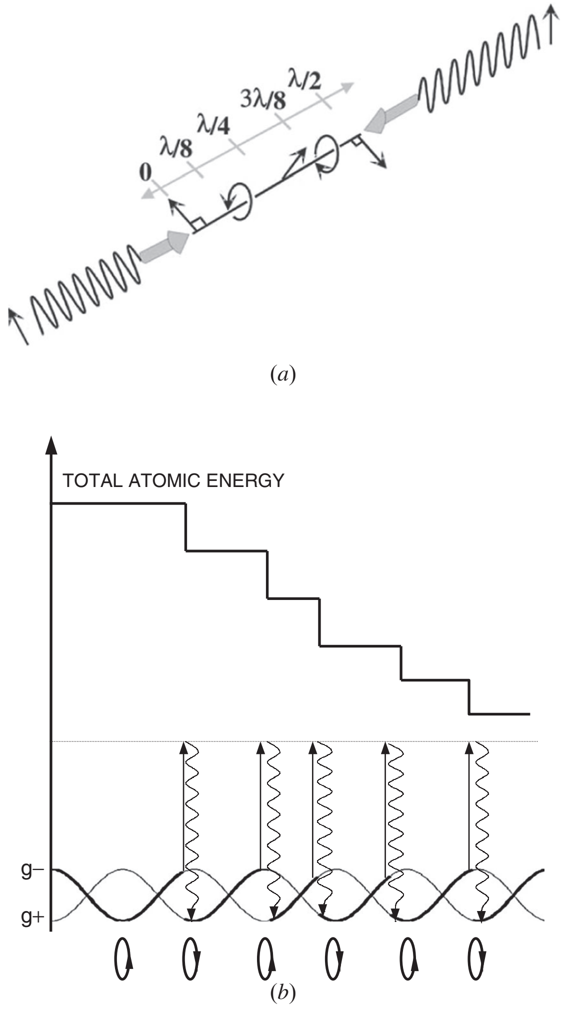
The first set of experiments with sodium [15] verified that it was possible to achieve temperatures close to the Doppler limit. However, further careful measurements revealed three major discrepancies with Doppler-cooling theory. The sensitivity to magnetic field was greater than predicted; the detuning dependence of the temperature was not followed and the killer blow—temperatures measured in the laboratory were an order of magnitude lower than expected [16]! This led to the realization that sub-Doppler-cooling mechanisms were responsible for the lower temperatures.

8.5 Sub-Doppler Cooling

Sub-Doppler cooling relies on two things which we previously neglected:

- Most atoms are not ideal two-state systems: The presence of multiple Zeeman levels in the hyperfine ground states introduces an additional time scale associated with optical pumping. The optical pumping time which characterizes the transfer of atoms between magnetic sub-levels is typically longer that the excited-state lifetime giving rise to the possibility of lower temperatures [13].
- The laser beams cannot be treated independently: The interference between beams leads to strong polarization gradients and, hence, to spatially dependent optical pumping rates.

In order to gain insight into the sub-Doppler cooling mechanisms, let us look at the example of a $J = 1/2$ ground state coupled to a $J = 3/2$ excited state. The two ground states interact differently with different polarizations, so if the polarization of the laser field is spatially dependent then the lowest-lying state is also spatially dependent. This is illustrated in Figure 8.3 for two equal-intensity plane waves with orthogonal polarizations. The ellipticity varies in space from linear (at 45°) at $z = 0$, to circular (driving σ^- transitions) at $z = \lambda/8$, to linear (at −45°) at $z = \lambda/4$, to circular (driving σ^+ transitions) at $z = 3\lambda/8$, and then the pattern repeats. The corresponding light shift of the two ground states is shown to the right. The crucial point is that optical pumping damps the system towards the lowest-lying ground state. This is equivalent to aligning the atomic dipole with the field. An atom initially at rest at $z = \lambda/8$ will be optically pumped into the $m_J = -1/2$ sub-level, g_-. If the atom starts to move it can end up in a higher energy state before optical pumping has occurred. For example, if the atom moves to the right at a speed corresponding to a quarter of a wavelength in an optical pumping time, the atom reaches the top of the potential hill, before being dumped into



(*a*)

(*b*)

FIGURE 8.3 (a) The polarization gradient light field produced by the interference of counter-propagating crossed linearly polarized laser beams. (b) The dressed energy levels in a standing-wave laser field. Spontaneous emission tends to occur when the atoms are at the top of the hill. Consequently, the total atomic energy gradually decreases until the atoms become trapped in the individual potential wells.

the g_+ sub-level. This sequence is repeated, and in each case, the atom is pumped from the top of a hill to the bottom of a valley. The process of always climbing hills transforms kinetic energy into potential energy which is radiated by the field. The effect is known as *Sisyphus cooling* after the character in Greek mythology condemned eternally to push a boulder up a hill.

The Sisyphus mechanism works well for atoms that are already cold and has a smaller capture velocity than conventional Doppler cooling. Faster atoms move through many cycles of the optical potential before being optically pumped and the effect averages out. This mechanism continues to operate until the atoms become cooled into the potential wells. Thus the final temperature is related to the well depth,

$$k_B T \sim \frac{\hbar\omega_R^2}{4|\Delta|} \tag{8.54}$$

which implies that, by increasing the detuning or reducing the intensity, one can lower the temperature. The fundamental limit is reached when the de Broglie wavelength of the atom becomes equal to the wavelength of light. This is known as the *recoil limit*,

$$\frac{1}{2}k_B T_{\text{rec}} = \frac{p^2}{2m} = \frac{\hbar^2 k^2}{2m}. \tag{8.55}$$

For ^{133}Cs, $T_{\text{rec}} = 100$ nK; however, in experiments temperatures of 1–2 μK are typical. Methods to cool below the recoil limit have been developed. Technically sub-recoil cooling is quite demanding and has only been realized by a handful of groups worldwide (for an overview, see, e.g., Ref. [8]).

8.6 Trapping of Cold Atoms

8.6.1 The Magneto-Optical Trap

Doppler cooling does not confine atoms; they can slowly diffuse out of the cooling region. To accumulate large numbers, it is necessary to make a trap. The most successful example is known as the *magneto-optical trap* or *MOT*. The MOT consists of a combination of magnetic fields and polarized light beams, such that an atom moving away from the centre is Zeeman-shifted into resonance with a beam that pushes it back. The simplest example, based on a $F = 0 \rightarrow F = 1$ transition, is illustrated in Figure 8.4. Beams 1 and 2 are left- and right-circularly polarized with angular momentum, $\ell = \pm 1$, along the z-axis, respectively. Note that we use the optics convention where left-circular means the E vector follows a left-handed corkscrew in space.

If the magnetic field $\boldsymbol{B}$ is parallel (anti-parallel) to ℓ, then σ^+ or $\Delta m_F = +1$(σ^- or $\Delta m_F = -1$) transitions will be driven.[2] This follows directly from the conservation of angular momentum: The initial state has a photon with angular momentum +1 plus an atom in the ground state $m_F = 0$. On absorption the photon disappears, and to conserve momentum, the atom must be transferred to an excited state with $m_F = +1$.

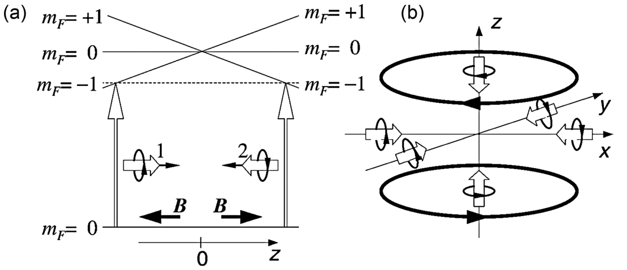


FIGURE 8.4 Principle of the MOT: The transitions and the relative directions of the beam angular momentum, ℓ, and B-field are shown in panel (a). Beams 1 and 2 excite σ^- transition for $z < 0$ and $z > 0$, respectively. The extension to three dimensions is illustrated on panel (b). A quadrupole field is produced by coils in an anti-Helmholtz configuration.

The magnetic interaction is of the form $U = g_F m_F \mu_B B$, where B is the magnitude of the magnetic field, g_F is the Landé g-factor and μ_B is the Bohr magneton. The MOT operates on a transition for which the excited state has a positive g_F; therefore, the state $m_F = +1$ always increases its energy in a magnetic field, whereas the state $m_F = -1$ decreases its energy. The MOT relies on a magnetic field of the form $B = b|z|$ which changes direction on either side of the origin. If the laser is red detuned, an atom at the origin will scatter equally from both beams and there is no net force. If the atom moves away from the origin, the Zeeman effect means that it scatters more photons from the beam which excites σ^- transitions. For $z < 0$, this is beam 1 and the atom is pushed back towards $z = 0$. Whereas for $z > 0$, it is beam 2, so again, the atom is pushed back. Writing the force on the atom as the sum of the two beams, and including the Zeeman shift in the detuning term, we have

$$F = F_1 - F_2 \tag{8.56}$$

where

$$F_{1,2} = \hbar k \frac{\Gamma}{2} \frac{I/I_s}{2I/I_s + 1 + 4(\Delta \mp b'z)^2/\Gamma^2}. \tag{8.57}$$

and $\hbar b' = g_F \mu_B b$. For small velocities, $k|v| < \Gamma$, and small displacements, $b'|z| < \Gamma$, we obtain the following expansion:

$$F = -\alpha v - \kappa z \tag{8.58}$$

where α is again the friction coefficient and $\kappa = b'\alpha/k$ is a 'spring constant'. The motion of atoms in an MOT is not dissimilar to that of an over-damped oscillator. The extension to three dimension, based on an anti-Helmholtz coil arrangement illustrated in Figure 8.4b, was first realized in 1987 [17].

The capture velocity of the MOT is enhanced by the Zeeman shift of the energy level. For typical experimental parameters, a typical capture velocity is a few tens of m s^{-1}. In early experiments, MOTs were loaded from a Zeeman-slowed atomic beam. A significant simplification for the heavier alkalis was the demonstration that a MOT could be loaded directly from the low-velocity tail of a room-temperature atomic vapour [18]. The vapour cell MOT is so easy to construct (see, e.g., Ref. [19]) and operate that it is now used in hundreds of laboratories worldwide.

---

[2] Note that care is needed with the labelling of a beam of light as, e.g., σ^+. It should be stressed that σ^+ is a label for a transition, not a beam. A σ^+ transition is of the form $\Delta m_F = m_{F_b} - m_{F_a} = +1$; a σ^- transition is of the form $\Delta m_F = -1$; a π-transition is of the form $\Delta m_F = 0$.

The MOT accumulates 10^6–10^{10} atoms in a cloud with a radius of around a millimetre at a temperature of around 10–100 μK. The phase-space density is much higher than in an atomic vapour but about six orders of magnitude below that needed for Bose–Einstein condensation. The minimum temperature is limited by spontaneous heating and the maximum density by trap loss due to excited-state collisions. For this reason, although the MOT and optical molasses are used as a starting point for Bose–Einstein condensation, the final step is made by evaporation cooling in a purely magnetic trap [4].

8.6.2 Magnetic and Optical Dipole Traps

Trapping neutral atoms is made difficult by their relatively (e.g. compared to ions) weak interaction with external fields. However, as laser cooling reduces the atomic kinetic energy by six orders of magnitude, trapping becomes straightforward. Two potentials are commonly used to trap and store laser-cooled atoms: the magnetic potential, $U = g_F m_F \mu_B B$, and the optical dipole potential (the gradient of which is the optical dipole force) given in equation (8.16).

Given that the magnetic moments of most atoms are of the order of the Bohr magneton and that laboratory fields of 100 G (0.01 T) are readily realized using current-carrying coils, one can produce a potential capable of trapping atoms with temperature of $\mu_B B/k_B$ which is of the order of milliKelvin, i.e. much larger than the temperature of laser-cooled atoms. A trap is realized by having a gradient of the potential, i.e. a spatial variation of the confining potential around the sample of cold atoms. By placing current-carrying wires close to a sample of cold atoms, it is possible to achieve large magnetic field gradients providing tight confinement. Atoms excited with circularly polarized laser beams and optically pumped into states with positive (negative) $g_F m_F$ are said to be weak (strong)-field seekers. As it is impossible to realize a magnetic field maximum in free space (Earnshaw's theorem), most static traps are designed to confine weak-field-seeking atoms. It is important to design a trap with a non-zero minimum; otherwise, non-adiabatic spin flips (Majorana transitions) can occur and atoms in the wrong internal magnetic state are ejected from the trap. Maxwell's equations place a constraint on the form of the potential achievable in magnetic traps and details of the geometries typically employed for guiding and trapping cold atoms can be found in [3,4].

The optical dipole potential or force encapsulated in equations (8.16) and (8.39) is also used to trap cold atoms. A confining potential is realized by having a spatial variation of the light intensity: this might be from a single-focused Gaussian profile laser beam; multiple crossed focused beams; standing waves (optical lattices) or evanescent waves formed when a beam is totally internally reflected at a prism–vacuum interface. If the laser frequency is lower than the atomic resonant frequency (red detuning), atoms are attracted to the most intense part of the beam, whereas atoms in a blue-detuned laser beam are repelled from regions of higher intensity. The trap depth is proportional to the laser power and inversely proportional to the detuning from resonance, whereas the spontaneous scattering rate is proportional to power divided by detuning squared (see Section 8.3.1); therefore, to reduce spontaneous scattering, a large detuning, the so-called far-off resonance optical traps, is often employed. The utility of the dipole force in guiding and trapping cold atoms is described further in [8,20].

8.7 Laser-Cooling Technology

The choice of atom for laser cooling is determined by the energy-level structure, which should provide a closed cooling cycle and the availability of narrow linewidth tunable laser sources at the desired wavelength. For example, early laser-cooling experiments were performed using sodium because rhodamine 6G dye lasers could be used to excite the cooling transition at 589 nm. Subsequently, the development of cheap and compact semiconductor diode lasers made laser cooling of rubidium and caesium very convenient. As mentioned earlier, the heavier alkalis have the additional advantage that the atoms can be trapped directly from a room temperature vapour. For the alkalis, transitions of the form $F \rightarrow F + 1$ are closed if the upper hyperfine state in the ground state is used as the lower level. An example, rubidium-85, is depicted in Figure 8.5a. The cooling laser is detuned between one-half and a few linewidths to the red of the $5^2S_{1/2}(F = 3) \rightarrow 5^2P_{3/2}(F = 4)$ transition. However, this is also far blue detuned from the $5^2S_{1/2}(F = 3) \rightarrow 5^2P_{3/2}(F = 3)$ transition leading to a small probability of the atom being excited into the $F = 3$ excited state and subsequently decaying into the wrong ground state. Consequently, a second laser is needed to re-pump atoms out of the lower ground state back into the cooling cycle. The re-pumper laser is tuned to resonance with the $5^2S_{1/2}(F = 2) \rightarrow 5^2P_{3/2}(F = 3)$ transition.

A second example, calcium-40, is depicted in Figure 8.5b. In this case, the cooling transition $4s^{2\,1}S_0 \rightarrow 4s4p\,^1P_1$ transition is at 422.8 nm. However, as there is no structure in the ground state, sub-Doppler cooling does not operate and the minimum temperature is limited to the Doppler limit of 0.83 mK. Again the cooling cycle is not strictly closed because the excited state can decay to the $3d4s\,^1D_2$ level and subsequently to the triplet state $4s4p\,^3P$ which is extremely long lived. This slow decay route effectively limits the lifetime of the MOT. One way to reduce the shelving in the triplet P state is to add a re-pumper resonant with the $4s4p\,^3P_1 \rightarrow 4s4d\,^1D_2$. The long-lived triplet P state can also be turned into an advantage as the narrow $4s^{2\,1}S_0 \rightarrow 4s4p\,^3P_1$ transition has a theoretical Doppler cooling limit of 8 nK! However, it would be far from trivial to get anywhere near this extremely low temperature because the recoil temperature (see Section 8.5) is 1.1 μK. In addition, the transition is so weak that the spontaneous force is not even strong enough to support against gravity. A list of atoms that have been laser-cooled with details of the cooling transition and re-pumper requirements is given in Ref. [8].

For efficient laser cooling, the laser linewidth must be less than the atomic linewidth. As typical cooling transition linewidths are of the order of 1–10 MHz, ultra-stable

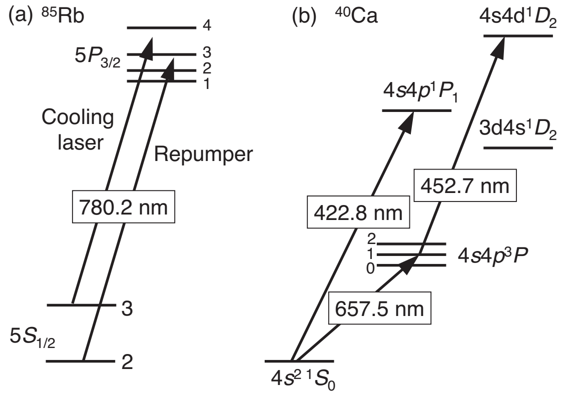


FIGURE 8.5 The energy levels involved in laser cooling of rubidium 85 (a) and calcium 40 (b).

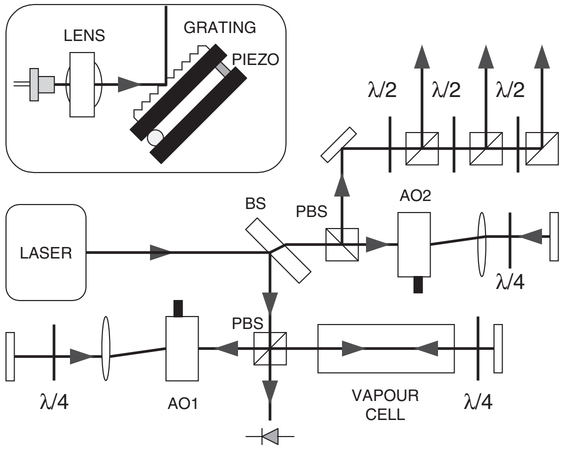


FIGURE 8.6 A possible optical layout for a standard laser-cooling experiment, where BS is a beamsplitter, PBS is a polarizing beamsplitter and λ/4and λ/2 are quarter- and half-wave plates. If the acousto-optic modulators, AO1 and AO2, are operated at frequencies ω_1 and ω_2, both on the +1 order, and the detector signal locks the light in the cell to the transition frequency ω_{ab}, then the light going to the experiment is detuned by $2(\omega_1 - \omega_2)$ to the red of the transition. A sequence of three half-wave plates and polarizing beamsplitter cubes is used to divide the light amongst the three orthogonal beam pairs. A schematic diagram of an extended cavity laser diode is shown inset.

narrow linewidth lasers are essential. For semiconductor diode lasers, the linewidth is reduced below 1 MHz by adding wavelength-selective feedback, for example, the first-order light from a diffraction grating. Extended cavity diode lasers employing a diffraction grating (see, e.g., Ref. [21]), as depicted in Figure 8.6(inset), have become the most widely used components in laser-cooling experiments. The ultra-high stability requirement also demands active control of the laser frequency. The feedback signal for active stabilization is usually provided by spectroscopy in a cell. A wide variety of spectroscopic techniques, such as saturation spectroscopy or polarization spectroscopy, are used. In Figure 8.6 we show a typical set-up where part of the laser output is split off to perform Doppler-free saturation spectroscopy in a vapour cell.

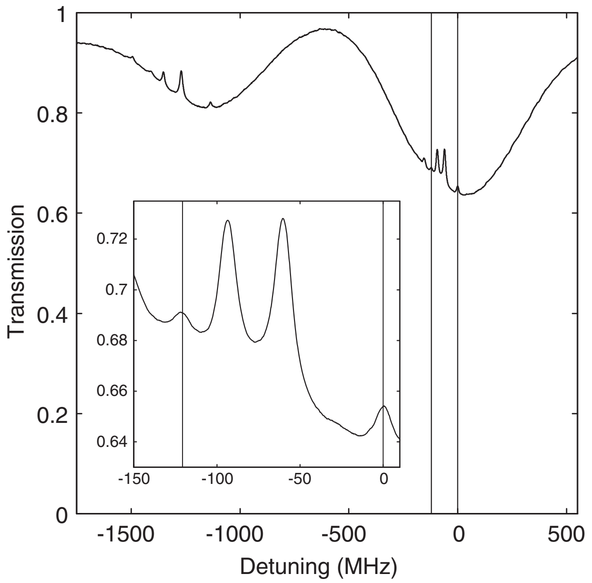


FIGURE 8.7 The absorption signal as a function of laser frequency for a rubidium vapour cell showing Doppler-free features including velocity-selective hyperfine pumping and saturated absorption. The $5^2S_{1/2}(F = 3) \rightarrow 5^2P_{3/2}(F = 3)$ and $(F = 4)$ transitions in rubidium-85 are indicated by vertical lines. The larger transmission peaks between these two lines (see inset) correspond to cross-over resonances. The small absorption line corresponds to the $5^2S_{1/2}(F = 2) \rightarrow 5^2P_{3/2}$ transition in rubidium-87.

The signal obtained for a rubidium vapour cell is shown in Figure 8.7. Typically, the laser is stabilized directly to the cooling transition or to the neighbouring cross-over resonance using a 'dither-lock'. Acousto-optic modulators are used to shift the frequency of the light going to the experiment to the red of the cooling transition.

8.8 Summary

In this chapter, we have attempted to give an overview of the theory underlying laser cooling and trapping and mention briefly some of the technology involved in laser-cooling experiments. The range of applications of laser cooling has grown enormously in a short time. It has become the standard technique in much of modern atomic physics and has led to the development of many new sub-fields including Bose–Einstein condensation of atomic vapours, atom interferometry, cold atom clocks, precision gravimetry, atom optics, optical lattices and atom lithography.

REFERENCES

1. Chu S., Cohen-Tannoudji C. and Phillips W. 1998 *Rev. Mod. Phys.* **70** 685.
2. Adams C. S., Sigel M. and Mlynek J. 1994 *Phys. Rep.* **240** 143.
3. Hinds E. A. and Hughes I. G. 1999 *J. Phys. D: Appl. Phys.* **32** R119.

4. Inguscio M., Stringari S. and Wieman C. (ed) 1999 Bose–Einstein condensation in atomic gases *Proc. Int. School of Physics Enrico Fermi* (Amsterdam: IOS Press).
5. Itano W. M., Bergquist J. C., Bollinger J. J. and Wineland D. J. 1995 *Phys. Scr. T* **59** 106.
6. GhoshP K. 1995 *Ion Traps (International Series of Monographs on Physics 90)* (Oxford: Clarendon).
7. Jessen P. S. and Deutsch I. H. 1996 Optical lattices *Adv. Atom. Mol. Opt. Phys.* **37** 95.
8. Adams C. S. and Riis E. 1997 *Prog. Quantum Electron.* **21** 1.
9. Metcalf H. J. and van der Straten P. 1999 *Laser Cooling and Trapping* (New York: Springer).
10. Einstein A. 1917 *Z. Physik.* **18** 121 (reprinted in English 1967 *Sources of Quantum Mechanics* ed B L van der Waerden (New York: Dover).
11. Ashkin A. 1997 *Proc. Natl. Acad. Sci.* **94** 4853.
12. Dalibard J. and Cohen Tannoudji C. 1989 *J. Opt. Soc. Am. B* **6** 2023.
13. Cohen-Tannoudji C. 1992 *Fundamental Systems in Quantum Optics* ed J. Dalibard, J-M. Raimond and J. Zinn-Justin (Amsterdam: North-Holland).
14. Prodan J. V., Phillips W. D. and Metcalf H. 1982 *Phys. Rev. Lett.* **49** 1149.
15. Chu S., Hollberg L., Bjorkholm J., Cable A. and Ashkin A. 1985 *Phys. Rev. Lett.* **55** 48.
16. Lett P. D., Watts R. N., Westbrook C. I., Phillips W. D., Gould P. L. and Metcalf H. J. 1988 *Phys.Rev. Lett.* **61** 169.
17. Raab E., Prentiss M., Cable A., Chu S. and Pritchard D. 1987 *Phys. Rev. Lett.* **59** 2631.
18. Monroe C., Swann W., Robinson H. and Wieman C. 1990 *Phys. Rev. Lett.* **65** 1571.
19. Wieman C., Flowers G. and Gilbert S. 1995 *Am. J. Phys.* **63** 317.
20. Grimm R., Weidemüller M. and Ovchinnikov Yu B. 2000 *Adv. At. Mol. Opt. Phys.* **42** 95.
21. MacAdam K. B., Steinbach A. and Wieman C. 1992 *Am. J. Phys.* **60** 1098.

FURTHER READING

Adams C. S. and Riis E. 1997 *Prog. Quantum Electron.* **21** 1.

Chu S., Cohen-Tannoudji C. and Phillips W. 1998 *Rev. Mod. Phys.* **70** 685.

Metcalf H. J. and van der Straten P. 1999 *Laser Cooling and Trapping* (New York: Springer).

9

Precision Timekeeping: Optical Atomic Clocks

Shimon Kolkowitz and Jun Ye

CONTENTS

9.1 Principles of Atomic Clocks 139
    9.1.1 Clock Stability 140
    9.1.2 Clock Accuracy 141
9.2 Ultra-stable Optical Cavities and Optical Frequency Combs 142
    9.2.1 Ultra-stable Optical Cavities 142
    9.2.2 Optical Frequency Combs 142
9.3 Optical Atomic Clocks 144
    9.3.1 Clock Interrogation Sequences 144
    9.3.2 Common Optical Clock Systematic Shifts 145
        9.3.2.1 Magnetic Fields 145
        9.3.2.2 Electric Fields 145
        9.3.2.3 Blackbody Radiation 145
        9.3.2.4 Doppler Shifts 146
        9.3.2.5 Gravitational Redshift 146
    9.3.3 Single-Ion Optical Clocks 147
        9.3.3.1 Ion Clock Operation 147
        9.3.3.2 Ion Clock Systematics 148
    9.3.4 Neutral Atom Optical Lattice Clocks 148
        9.3.4.1 Optical Lattice Clock Operation 148
        9.3.4.2 Optical Lattice Clock Systematics 150
9.4 Outlook and Future Directions 151
    9.4.1 Next-Generation Clocks 151
        9.4.1.1 3D Optical Lattice Clocks 151
        9.4.1.2 Cryogenic Optical Clocks 152
        9.4.1.3 Superradiant Optical Clocks 152
        9.4.1.4 Entangled Clocks 152
        9.4.1.5 Exotic Atomic Clocks 152
    9.4.2 Emerging Applications of Optical Clocks 152
        9.4.2.1 SI Unit Definitions, the World Clock, Navigation, and Geodesy 152
        9.4.2.2 Searches for Variations of Fundamental Constants, Dark Matter, and Gravitational Waves 153
        9.4.2.3 Many-Body Quantum Physics 153
References 153
Further Reading 156

9.1 Principles of Atomic Clocks

Clocks are the metrological tools used to measure the passage of time in seconds, the SI unit for time. All precise clocks do this by counting the cycles of a periodic physical phenomenon, such as the oscillations of a pendulum, as in a grandfather clock, or the vibrations of a piezoelectric tuning fork, as used in modern electronics. Because these periods ideally occur regularly, the number of counted periods is a measure of the amount of time that has passed. However, because no two man-made oscillators are ever identical, and because the frequency of their oscillations is coupled to their changing external environments, two oscillators will inevitably drift away from each other, limiting their stability, and both oscillators will always have a finite error relative to the agreed upon standard definition of the second, limiting their accuracy.

In contrast, all atoms of the same elemental and isotopic species are indistinguishable. Atomic clocks take advantage of this to achieve greater levels of both stability and accuracy by referencing the oscillations of a man-made "local oscillator," such as a piezoelectric tuning fork or the electro-magnetic modes of a microwave or optical cavity, to the resonance frequency of a "clock transition" between two internal energy levels in an atom of choice. This also enables the definition of the SI second with respect to the frequency of this transition, a constant of nature that, at least as far as we know, is the same in every inertial frame of reference (a principle that is foundational to the theories of both special and general relativity).

However, whilst two atoms placed in identical conditions will have the same internal structure, their energy levels can be shifted with respect to each other by external perturbations such as electric and magnetic fields. The achievable accuracy and stability of a particular atomic clock is therefore determined by three related factors: (i) how well the effects of external perturbations on the energy levels of the two clock states are understood, limiting accuracy; (ii) how precisely and rapidly deviations in frequency between the local oscillator and the atomic frequency reference can be measured and corrected, limiting stability; and (iii) how well the external environment can be measured and controlled, which limits the reproducibility of the clock, and therefore both its accuracy and long-term stability. In the following subsections, we elaborate on these factors and their roles in clock performance, including the relevant figures of merit.

9.1.1 Clock Stability

The fractional frequency instability (often simply called the stability) of a clock, $\sigma_y(\tau)$, is a unitless number that parameterizes how precisely deviations in frequency between the local oscillator and the atomic frequency reference can be measured and corrected over a measurement of total duration τ. The smaller $\sigma_y(\tau)$, the more stable, or precise, the clock. One way to think about the stability of a clock is as the consistency with which the clock keeps time over some period of time. As a simple example, consider two clocks, *A* and *B*, both of which nominally oscillate at 1 GHz. We assume that clock *B* is much more stable than clock *A* and can therefore serve as a reference clock. If over the course of many comparisons of duration 1 s clock *A* is measured to have a standard deviation in frequency of 5 Hz with respect to clock *B*, then the stability of clock *A* is $\sigma_y(1\,\text{s}) = 5\ \text{Hz}/1\ \text{GHz} = 5\times10^{-9}$. If clock *A* was then used to keep time, over the course of 1 s on average it would lose or gain ±5 nanoseconds. This simple example highlights the difficulty in determining the stability of a very stable clock, in that it can only be measured through comparisons with other clocks of comparable stability.

At short timescales (on the order of seconds to hours) $\sigma_y(\tau)$ is typically limited by noise that obscures the measurement of the offset of the local oscillator frequency from the frequency of the atomic reference. Therefore, the less noisy the local oscillator and the atoms are, the smaller $\sigma_y(\tau)$ will be and the more consistently the clock will keep time from second to second. Often the noise at these timescales has a flat frequency spectrum and therefore can be averaged down. The stability of an atomic clock therefore typically improves with averaging time τ for hours of measurement.

We now consider the stability of an atomic clock more quantitatively. As described in the previous section and shown in Figure 9.1, an optical atomic clock operates by locking the frequency of a drifting local oscillator to that of a narrow atomic transition. This is typically done by preparing *N* atoms into the ground clock state $|g\rangle$ and then driving the atoms with radiation at the frequency of the local oscillator and measuring the subsequent atomic populations $N_{g,e}$ in states $|g\rangle$ and $|e\rangle$ using a strong optical cycling transition (Nagourney et al. 1986), which enables high-fidelity projective single-shot measurement of each atom's state. With a judicious choice of measurement sequence, driving strength, and detuning (see Section 9.3.1), the excitation fraction $P_e = N_e/(N_e+N_g)$ will be linearly proportional to the detuning from resonance Δ and can be used as a discriminator signal to feed back on the local oscillator and keep it locked to the atomic reference. The statistical errors in the measurement of Δ introduced by a combination of fundamental and technical noise sources will limit the stability $\sigma_y(\tau)$ of the clock.

The measurement of P_e, and therefore of Δ, is fundamentally limited by the quantum noise associated with the projective measurement of each atom's state into either $|g\rangle$ and $|e\rangle$. In the absence of any other sources of noise and for a clock interrogation time of length *T* with no dead time between measurements, the clock stability will be given by

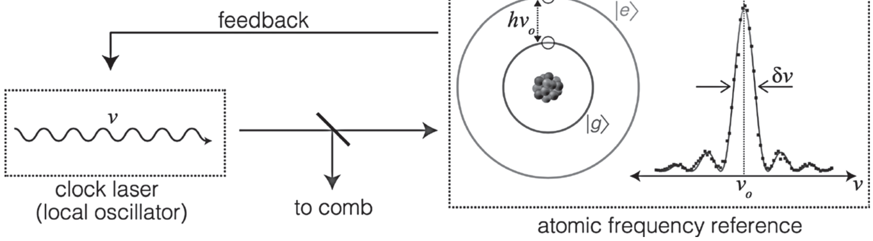


FIGURE 9.1 Diagram of optical atomic clock operation. In an optical atomic clock, a stable laser at frequency ν serves as the local oscillator. It is used to interrogate atoms prepared in the ground state $|g\rangle$ of an ultra-narrow optical atomic transition whose frequency ν_o and linewidth $\delta\nu$ serve as the frequency reference. The number of atoms in the excited state $|e\rangle$ after the clock interrogation is used to feed back on the clock laser frequency and stabilize it at or near $\nu = \nu_o$. The frequency-stabilized clock laser is sent to an optical frequency comb, which is used to convert the optical frequency either to a microwave clock signal or to another optical frequency. The data were taken with a strontium optical lattice clock in the Ye group at JILA and is an example of Rabi spectroscopy of an optical clock transition (see Section 9.3.1).

$$\sigma_y(\tau) = \frac{1}{2\pi\nu_0 T}\sqrt{\frac{T}{N\tau}}, \tag{9.1}$$

where ν_o is the frequency of the transition and N is the number of atoms in the clock (Campbell et al. 2017). This equation is quite easy to understand. The first term is the inverse number of frequency reference cycles that are counted in a single measurement (in inverse radians). The more cycles that can be counted in a single measurement, the more information is gained about the relative detuning of the oscillator from the frequency reference. The second term is the square root of the inverse total number of single atom measurements performed in time τ, which can be interpreted as the improvement of signal-to-noise ratio with more averages. From this equation, we can see that for a quantum projection noise-limited atomic clock, the stability will improve with increasing averaging time, scaling with $\tau^{-1/2.}$

Equation (9.1) offers a number of insights into the design of the ideal atomic clock. First, the scaling with $1/\sqrt{N}$ indicates that interrogating a large number of atoms simultaneously is advantageous. Second, the overall factor of $\sqrt{T}$ in the denominator indicates that a long-lived atomic transition is desirable, as the lifetime of the transition imposes an ultimate limit on the length of a single-clock measurement. Fortunately, a number of candidate atoms exhibit appropriately named "clock transitions" which are highly insensitive to external perturbations and are very long lived. To take full advantage of these transitions, a local oscillator with as long of a coherence time as possible should also be used. However, the interrogation time can also be limited by other sources of dephasing. The atoms should therefore be well isolated from their surrounding environment and from interactions with each other to reduce inhomogeneous broadening of the atomic transition, and should be as cold as possible to reduce Doppler broadening. As the temperature limits of laser cooling are insufficient to narrow the Doppler linewidth down to the natural linewidths of optical clock transitions, the atoms should be trapped in a confining potential to further decouple the clock transition from the finite atomic temperature, and to increase the interrogation time T to the transition lifetime $T_e = 1/2\pi\delta\nu$, where $\delta\nu$ is the clock transition linewidth (Ludlow et al. 2015, Poli et al. 2014).

Finally, and perhaps most importantly in the context of this chapter, the factor of ν_o in the denominator explains the advantage of moving from microwave atomic clocks, with transition frequencies of $\nu_o \approx 10^{10}$ Hz, to optical atomic clocks, with frequencies of $\nu_o \approx 10^{15}$ Hz. This can intuitively be understood as the advantage of counting seconds instead of counting days when trying to accurately measure the passage of time (Diddams et al. 2004). Simply put, the more cycles that can be counted in a single measurement, the better the measurement will be. All other factors remaining equal, the use of optical frequency transitions offers orders of magnitude gains in clock stability. Of course, this is much easier said than done. At present there are no electronics fast enough to count the oscillations of optical frequency electromagnetic radiation, so it was not until the start of the 21st century and the advent of the optical frequency comb, which enables the transfer of stability from an optical local oscillator to the microwave domain (see Section 9.2.2), that optical clocks became a practical reality.

As the measurement and local oscillator noise are averaged down, at longer timescales (typically on the order of hours to days), $\sigma_y(\tau)$ can become limited by uncontrolled low-frequency fluctuations of the surrounding environment, and equation (9.1) will no longer be valid. This results in slowly varying systematic shifts that limit the reproducibility of the clock frequency measurement, and at these time scales $\sigma_y(\tau)$ will either bottom out or even begin to grow with τ (Ludlow et al. 2015, Poli et al. 2014). Of course, the reproducibility of the clock frequency represents a limit not only on the day-to-day stability of a clock but also on the level of agreement it can be expected to have with a different clock or with the accepted definition of the second, which is parameterized by the clock accuracy.

9.1.2 Clock Accuracy

In addition to the statistical errors described in Section 9.1.1, the accuracy of an atomic clock is limited by systematic errors associated with shifts of the atomic clock transition frequency due to the surrounding environment. The extent to which systematic shifts can be anticipated, characterized, and controlled determines the level of accuracy at which the true frequency of the atomic transition can be determined and, therefore, the level of agreement that can be expected between the frequencies of two different clocks at long measurement times (Ludlow et al. 2015, Poli et al. 2014).

While the frequency of a particular atomic transition is a constant of nature, it is not a fundamental constant, and we do not presently have the tools to predict the frequency of a particular clock transition to anywhere near the accuracy with which we can measure it. The goal is therefore to anticipate, evaluate, and control all possible systematic shifts of a clock-transition frequency through a combination of theory and experiment, and to then verify the consistency of these evaluations through comparisons between clocks. Because the accuracy of optical clocks now exceeds the fractional precision with which the SI second is defined (Ludlow et al. 2015), optical clock comparisons now consist of precise measurements of the frequency ratios between optical transitions, which are absolute constants that are independent of unit definitions, and can be measured at the level of precision set by the constituent clocks.

Common sources of systematic shifts of clock transition frequencies include electric and magnetic fields, interactions between atoms, and motion of the atoms with respect to the lab frame which results in Doppler shifts of the atomic transition frequency. In addition, relative motion and differences in altitude on Earth will result in relativistic shifts that will appear in comparisons between clocks. It is not always necessary or even desirable to eliminate such shifts, but it is critical that they be known to the desired level of accuracy and included in an overall accounting of all the systematic effects. In Section 9.3.2 we summarize some of the major systematics for optical clocks. However, the dominant systematics can vary considerably for each atomic species and experimental apparatus, and we briefly discuss some of the systematics that are specific to ion and lattice clocks in Sections 9.3.3 and 9.3.4, respectively. In addition, as clock accuracy continues to improve through better understanding and control of the current limiting systematic effects, previously insignificant or unforeseen effects will undoubtedly emerge. While this represents a challenge to developing ever more accurate clocks, it also represents an opportunity, as many of these effects are interesting and worthy of study in their own right, as we discuss in Section 9.4.2.

Despite the distinctions we have drawn between accuracy and stability, they remain intimately connected. As discussed in Section 9.1.1, if the environment is not well controlled and the systematic shifts drift with time the reproducibility of the clock will be limited, and the long-term stability may be worse than the short-term stability, as with hydrogen masers or quartz oscillators. Similarly, the stability of a clock often determines how quickly a particular systematic shift can be evaluated through averaging. As a result, clock stability and accuracy have historically progressed hand in hand.

9.2 Ultra-stable Optical Cavities and Optical Frequency Combs

9.2.1 Ultra-stable Optical Cavities

In an optical atomic clock the local oscillator must oscillate at an optical frequency resonant with the atomic transition, which naturally suggests the use of a laser. However, as discussed in Section 9.1.1, in order to achieve high levels of clock stability the local oscillator should have a linewidth comparable to the $\sim 1-1000$ mHz atomic clock transition linewidth, whilst free-running lasers typically have linewidths many orders of magnitude broader than this. Furthermore, in current optical clocks noise on the clock laser is actually even more detrimental due to the "Dick effect," in which the finite dead time required to cool and reload atoms in each measurement cycle leads to laser-noise-limited clock stabilities well above the quantum projection noise limit given by equation (9.1). The Dick effect can be understood as an aliasing down of higher frequency laser noise due to the stroboscopic measurements with the atoms (Santarelli et al. 1998, Westergaard et al. 2010) and is currently the limiting factor for optical lattice clock stabilities (Schioppo et al. 2017). The desire to improve achievable clock stabilities has therefore motivated a concerted research effort into the generation of ultra-stable laser light.

While novel techniques for laser stabilization such as spectral hole burning in rare earth-doped crystals and hybrid atom-cavity geometries are currently being explored (Cook et al. 2015, Christensen et al. 2015), so far the most successful and widely adopted laser stabilization technique is to lock the laser to a passive, ultra-stable, high-finesse, two-mirror optical cavity (Matei et al. 2017) using the Pound-Drever-Hall locking technique (Drever et al. 1983), which is described in Chapter [Laser Stabilization for Precision Measurements]. The linewidth of the laser is then determined by the stability of the cavity resonance frequency, which in turn depends on the cavity length. To stabilize the cavity length, the two mirrors are mounted on either end of a transparent, solid spacer. The spacer is shaped and mounted to minimize length changes due to accelerations of the cavity from acoustic and mechanical vibrations of the surrounding environment (Swallows et al. 2012, Nazarova et al. 2006, Millo et al. 2009). The spacer material is selected to have a low coefficient of thermal expansion at the operating temperature, and the cavity is kept in a temperature-stabilized, thermally shielded vacuum chamber. The end result is that the frequency stability of these cavities at relevant clock interrogation times is no longer limited by their coupling to the outside environment, but rather by the thermal fluctuations of the spacer material and mirror coatings (Bishof et al. 2013, Ludlow et al. 2007, Webster et al. 2008, Jiang et al. 2011). These thermal fluctuations can be reduced by moving to cryogenic temperatures in vibration-isolated cryostats and by using cavities with rigid single-crystal spacers (Kessler et al. 2012), Matei et al. 2017), Zhang et al. 2017)) and crystalline mirror coatings (Cole et al. 2013). These efforts have recently culminated in the demonstration of sub-10 mHz laser linewidths and corresponding laser stabilities of $\sigma_y = 4\times10^{-17}$ at 1 s (Matei et al. 2017), as shown in Figure 9.2.

9.2.2 Optical Frequency Combs

The ultra-stable optical cavities described in Section 9.2.1, combined with the techniques for stabilizing an optical local oscillator to a narrow optical clock transition described in Section 9.3, enable the generation of remarkably stable lasers with a known absolute frequency relative to the natural atomic transition. However, the utility of this laser would be quite limited without the ability to compare it to other optical frequencies or to reference its frequency down to the microwave regime so that it can be counted by modern electronics and compared to microwave atomic frequency standards. Fortunately, this is made possible by the optical frequency comb.

To understand how a frequency comb can be used to transfer the stability of an optical-clock-stabilized local oscillator at frequency f_L down to microwave frequencies, consider a mode-locked laser (see Chapter [Mode-Locked Lasers] for details) with a repetition rate of f_{rep} (typically ~100–1000 MHz), and a carrier-envelope offset frequency f_0 as shown in Figure 9.3. The nth comb tooth will have a frequency given by $f_n = f_0 + nf_{\text{rep}}$. When the mode-locked laser is free running, f_0 can take any value between $0 - f_{\text{rep}}$ and will drift over time. However, if the mode-locked laser is sent through a non-linear optical element such as a photonic crystal fibre to generate a harmonic spanning spectrum, f_0 can be stabilized using a $f-2f$ interferometer. The nth comb tooth is frequency-doubled using second-harmonic generation and interfered with the $2n$th comb tooth at frequency f_{2n}, to produce a beat note at $2f_n - f_{2n} = 2\left(f_0 + nf_{\text{rep}}\right) - \left(f_0 + 2nf_{\text{rep}}\right) = f_0$, which can then be stabilized to a microwave reference, such as a GPS-trained 10 MHz oscillator or a caesium or rubidium vapour cell clock.

If an optical clock laser is then interfered with the mth comb tooth with frequency f_m it will produce a beat note at $f_m - f_L = f_0 + mf_{\text{rep}} - f_L$. This beat note can then be stabilized by feeding back on f_{rep}, such that the microwave frequency f_{rep} now has the stability of the stabilized clock laser. While the frequency f_{rep} will also depend on f_0, its contribution is divided down by a factor of m where $m \approx 10^{14}$ Hz/100 MHz $= 1\times10^6$. As a result the stability of the microwave standard is divided down by a large enough factor to make its contribution

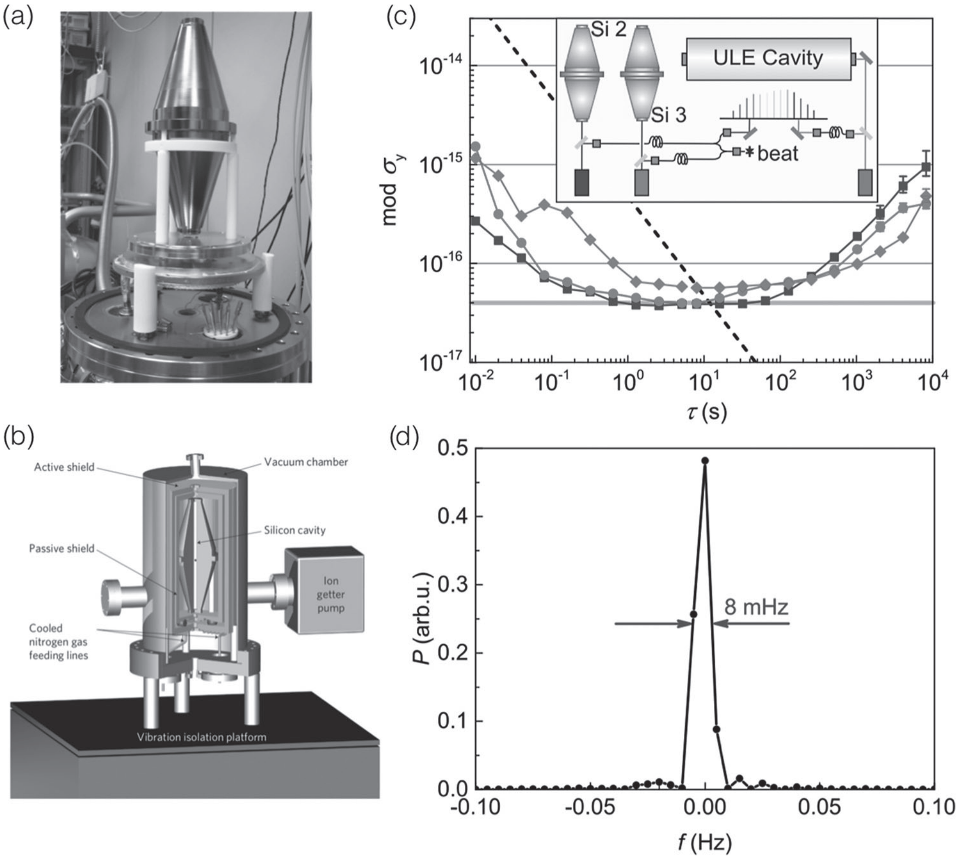


FIGURE 9.2 Ultra-stable optical cavities. (a) One of the two single-crystal silicon optical cavities from the JILA-PTB ultra-stable cavity collaboration, which has been machined and mounted to minimize susceptibility to accelerations. (b) Schematic of the single-crystal silicon cavity (a) in a nitrogen-gas-based vibration-isolated cryostat, including the vacuum chamber and two heat shields. Figure published in Kessler et al. (2012). (c) A comparison between the two cryogenic single-crystal silicon optical cavities and a ULE cavity, demonstrating silicon cavity stabilities of $\sigma_y = 4\times10^{-17}$ at 1 s. Data and figure published in Matei et al. (2017). (d) A measured sub-10 mHz linewidth beat note between the two single-crystal silicon optical cavities. (Data and figure published in Matei et al. 2017.)

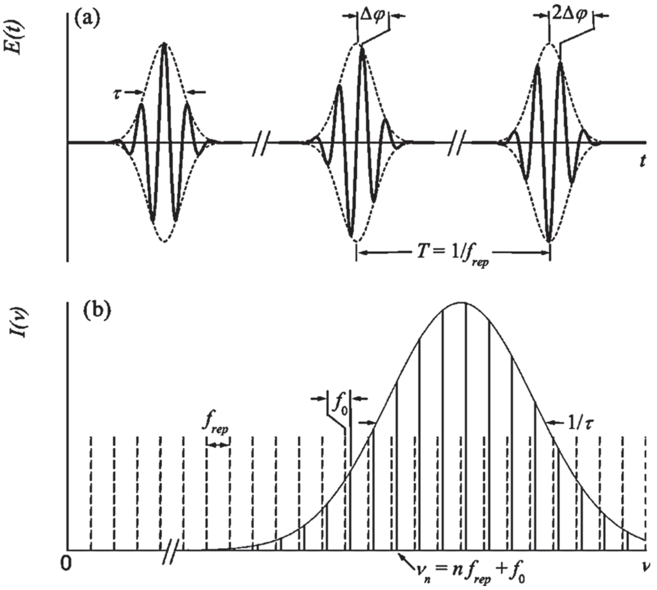


FIGURE 9.3 The optical frequency comb. (a) In the time domain, a frequency comb consists of a train of periodically spaced femtosecond laser pulses with well-defined carrier phase relationships between each pulse. (b) In the frequency domain, this corresponds to a broad spectrum of equally spaced, sharp "comb teeth." The frequency of each tooth is uniquely determined by f_{rep} and the phase offset between subsequent pulses $\Delta\phi$, which is related to f_0 by $2\pi f_0 = f_{\text{ref}}\Delta\phi$. (Figure from Cundiff & Ye 2003.)

negligible, and the result is a microwave oscillator referenced to the frequency of the clock transition, and with the same fractional frequency stability as the clock laser itself.

With the frequency comb serving as the bridge, optical clocks can be used as ultra-stable microwave frequency references, and the frequency of optical atomic transitions can be measured relative to the current SI unit definition of the second based on the caesium hyperfine transition (Udem et al. 2001, Hoyt et al. 2005, Stenger et al. 2001, Margolis et al. 2003, Campbell et al. 2008). Furthermore, the broad spectral span of the comb teeth enables direct frequency ratio comparisons between optical clocks that make use of atomic clock transitions at different optical frequencies (Rosenband et al. 2008, Yamanaka et al. 2015, Nemitz et al. 2016).

9.3 Optical Atomic Clocks

In order to achieve high levels of stability, optical clocks rely on atoms with narrow linewidth optical transitions to maximize the number of cycles in a single measurement, corresponding to minimizing $1/\nu_o T$ in equation (9.1). These are typically either forbidden optical quadrupole transitions in alkali-like ions, or narrow inter-combination transitions in alkaline earth and alkaline-earth-like atoms and ions. Each particular species has its own set of advantages and disadvantages, including the clock transition linewidth, susceptibility to systematic shifts, and the existence and convenience of transitions for laser cooling, state preparation, and optical state read-out.

As one example, in alkaline-earth(-like) atoms, the doubly forbidden $^1S_0$ to $^3P_0$ transition is weakly allowed by hyperfine state mixing in fermionic isotopes with non-zero nuclear spin but is completely forbidden in bosonic isotopes with no nuclear spin. Optical clocks that make use of the bosonic isotopes therefore must artificially introduce state mixing by applying strong magnetic fields or driving additional transitions out of the clock states, which introduce additional systematic shifts. Optical clocks that make use of the fermionic isotopes can avoid these additional environmental perturbations, but the additional hyperfine states make laser cooling more complex and also contribute additional complexity to the evaluation of ac Stark shifts through their coupling to optical polarization (see Section 9.3.4.2).

In addition to the variety of atomic species and transitions to choose from, there are presently also two different leading realizations of optical atomic clocks: single ion clocks and optical lattice clocks. In both cases, the atoms/ions are tightly confined in a trapping potential to isolate the internal clock transition from the motional degrees of freedom. In the following subsections, we describe some of the measurement sequences and systematic shifts that are common to both kinds of optical clock. We then discuss some of the experimental protocols, techniques, and systematics that are specific to each approach.

9.3.1 Clock Interrogation Sequences

To lock the frequency of an ultra-stable clock laser to a narrow atomic clock transition in either an ion or optical lattice clock an appropriate pulse sequence must be applied to convert the resulting clock state population into a discriminator for the detuning of the clock laser from resonance. For this a Ramsey or Rabi spectroscopy sequence can be employed. To understand these sequences, we restrict ourselves to a two-level system defined by the two-clock states, which we label $|g\rangle$ and $|e\rangle$ for the ground and excited states, respectively. Both sequences begin with the atoms prepared in $|g\rangle$.

In the case of the Rabi spectroscopy sequence a single pulse of the clock laser light of length $T_{\text{Rabi}} = \pi/\Omega$ is applied to the atoms, where Ω is the Rabi frequency of the clock light. This is known as a π-pulse, meaning that if the clock laser is exactly on resonance with the atomic transition the atoms will be undergo π radians of a Rabi cycle, and following the pulse all of the atoms will have been coherently transferred to the $|e\rangle$ state. If the detuning of the clock laser from resonance Δ is scanned over the transition and the resulting excited state fraction $P_e(\Delta)$ is measured, the result will be a lineshape of the form $P_e(\Delta) = \frac{\Omega^2}{\Delta^2 + \Omega^2}\sin\left(\frac{\pi\sqrt{\Delta^2+\Omega^2}}{2\Omega}\right)^2$ as shown in Figure 9.6c. The clock laser can then be locked to the atomic resonance by fixing the detuning Δ at a point halfway up the slope of the resonance where $P = 0.5$. In the absence of other broadening mechanisms, the width of the Rabi line will be Fourier limited and will scale inversely with T_{Rabi}. Ideally, T_{Rabi} would therefore be set to the lifetime of clock transition; although in practice, it is often limited by the linewidth of the laser, or by other broadening mechanisms.

In the case of Ramsey spectroscopy, the atoms are first driven with a strong $\pi/2$ clock pulse, which places them in a coherent superposition of the two spin states, $|\Psi\rangle = \frac{1}{\sqrt{2}\left(|g\rangle + |e\rangle\right)}$. Left in this state for interrogation time T, the superposition will acquire a relative phase $\phi = \Delta T$, where Δ is the detuning between the laser and the transition angular frequency $\omega_o = 2\pi\nu_o$. With a second strong $\pi/2$ clock pulse at the end of the sequence, the phase ϕ is then converted into a population difference between the two clock states, which can be measured by driving the atoms on a cycling transition out of one of the clock states and collecting the emitted photons. If the Rabi frequency is much larger than the detuning Δ, the fraction of the atoms in $|e\rangle$ at the end of the Ramsey sequence will be given by $P_e = \cos\left(\frac{\Delta T}{2}\right)^2$, and sweeping Δ will produce oscillating Ramsey fringes (as shown in Figure 9.6d), which can be used to generate an error signal and lock Δ to a fixed value. Once again the stability will improve with longer T, and ideally T would correspond to the lifetime of the clock transition. In fact, the clock stability achievable with an ideal Ramsey sequence of length T will be essentially the same as that of an ideal Rabi sequence of length T_{Rabi}. The choice of which sequence to use is therefore dependent on technical considerations specific to each particular clock, such as atomic interaction effects and clock laser ac Stark shifts.

9.3.2 Common Optical Clock Systematic Shifts

The primary systematic effects that are common to both ion and optical lattice clocks are shifts due to electromagnetic fields at the location of the atom, and relativistic shifts arising from atomic motion and gravity. We now briefly discuss each of these effects and the techniques employed to characterize and control them.

9.3.2.1 Magnetic Fields

Static magnetic fields are often intentionally applied to the atoms in an optical clock to define a quantization axis for better internal state control and to lift the degeneracy of hyperfine states in atoms with non-zero nuclear spin. In addition, stray magnetic fields due to the Earth's magnetic field or equipment in the lab may be present at the location of the atoms. The effect of magnetic fields on the clock transition frequency must therefore be accounted for. In general, the ground and excited clock states will have different magnetic moments, and the resulting shift of a particular clock transition in a small magnetic field $\vec{B}$ can be written as an expansion in $\left|\vec{B}\right|$,

$$\Delta\nu_B = C_1\left|\vec{B}\right| + C_2\left|\vec{B}\right|^2 + \cdots, \qquad (9.2)$$

where C_i is the ith order magnetic field coupling coefficient for the transition (Ludlow et al. 2015). Typically, only the first two terms need to be considered. Because C_1 is proportional to the angular momentum quantum number m_f, the linear first-order term can either be made zero by using a transition between two clock states with $m_f = 0$ in isotopes with integer angular momentum, or can be cancelled by measuring the transition frequency of two transitions with symmetric shifts, such as between two clock states with $m_f = +9/2$ and between two clock states with $m_f = -9/2$, and averaging them to cancel the shift, as is commonly done in 87Sr optical lattice clocks. The second-order shift can be accounted for by first characterizing C_2 using large magnetic fields and then periodically measuring $\left|\vec{B}\right|$ during clock operation using states with a non-zero C_1.

9.3.2.2 Electric Fields

Electric fields can shift the frequency of the clock transition through the dc Stark and ac Stark effects. While it is unusual to intentionally apply non-zero static electric fields to clock atoms, stray static electric fields may be present due to charge build-up on nearby surfaces. Due to their charge, ions will always move to sit at zero dc electric field, so in ion clocks, there is no dc Stark shift. However, stray electric fields can displace the ion from the null point of the RF fields used to trap the ion, resulting in excess micromotion of the ion and a quadratic Stark shift from the root mean square amplitude of the oscillating-trapping fields.

In contrast to ions, neutral atoms in optical lattice clocks are happy to sit at finite electric fields. The lack of an electric dipole moment in either clock state means that the linear dc Stark shift is negligible, but the second-order quadratic Stark shift arising from stray static electric fields can be significant (Nicholson et al. 2015). The traditional approach in optical lattice clocks has been to carefully characterize and even cancel the dc electric field environment by measuring the Stark shift as additional electric fields are applied via external electrodes (Bloom et al. 2014, Nicholson et al. 2015). This approach has been sufficient to reach dc Stark shift uncertainties at the 10^{-18} level, and recently dc Stark shift uncertainties at the 10^{-20} level were demonstrated in a ytterbium optical lattice clock through a combination of in-vacuum Faraday-shielding and in-vacuum electrodes (Beloy et al. 2018).

Many of the most significant systematics for current optical clocks are ac Stark shift effects from oscillating electric fields. Oscillating electric fields are always present in both ion clocks and optical lattice clocks, due to blackbody radiation (BBR) from the surrounding environment, as well as the clock laser light used to probe the clock transition, which even when on resonance can itself shift the relative energies of the ground and excited clock states due to the existence of other transitions out of both states. In ion traps, there are also the RF fields used to confine the ions, and in the case of quantum logic-based ion clocks (see Section 9.3.3), there are sympathetic cooling lasers. In optical lattice clocks there is also the optical lattice itself, and the frequency and polarization of the lattice must be carefully selected and controlled to minimize the differential ac Stark shift of the clock transition (see Section 9.3.4).

9.3.2.3 Blackbody Radiation

While all of the various ac Stark shifts are important, perhaps the most common limiting systematic for current generation optical clocks is the BBR shift. While the other ac fields are intentionally applied and can be varied to provide a "lever arm" with which to characterize their impact (Bloom et al. 2014), it is difficult to vary the temperature over a large range for room temperature clocks, and the thermal environment is often inhomogeneous and varies with time.

At room temperature the energy density of the BBR spectrum is concentrated at much lower frequencies than the electric dipole-allowed transitions out of the clock states, and so can be approximated as a dc electric field. A higher order "dynamic" correction term is introduced to account for the deviations from this approximation due to the specific transitions out of the clock state that couple most strongly to the blackbody spectrum. The resulting temperature-dependent shift in the clock transition frequency is given by

$$\Delta\nu_{\mathrm{BBR}} = \frac{\Delta\alpha_{\mathrm{s}}\left\langle E^2(T)\right\rangle}{2h}\left(1+\eta\left(T^2\right)\right) \qquad (9.3)$$

where $\Delta\alpha_{\mathrm{s}}$ is the differential static scalar polarizability and $\eta(T^2)$ is the dynamic correction term for the clock states, h is Planck's constant, and the temperature dependence of the mean-squared amplitude of the blackbody spectrum can be found by integrating Planck's law, $\left\langle E^2(T)\right\rangle\left(831.9\ \mathrm{V\,m^{-1}}\right)^2\left(T/300\ \mathrm{K}\right)^4$ (Nicholson et al. 2015). The uncertainty in the BBR shift for a specific atomic clock can either come from uncertainty in the effective temperature at the atoms, or in the uncertainty at which $\Delta\alpha_{\mathrm{s}}$ and $\eta(T^2)$, which vary considerably for different

TABLE 9.1

BBR Coefficients for the Clock Transitions in Commonly Used Optical Clock Species, and the Corresponding Frequency Shift and Fractional Frequency Shift at 300 K

Atom/Ion	**$\Delta\alpha_s$ (10−41 J m2 V−2)**	**$\eta(T^2)$**	**$\Delta\nu_{BBR}$, 300 K (Hz)**	**$\Delta\nu_{BBR}/\nu_o$, 300 K (10^{-16})**
Al+	0.82(8)	$<10^{-4}$	−0.0043	−0.0038
Ca+	−73.0(1.0)		0.380	9.3
Sr+	−47.938(71)	−0.00951(15)	0.248	5.6
Yb+ *E*2	69(14)		−0.36	−5.2
Yb+ *E*3	13(6)		−0.067	−0.97
Hg+	15		−0.079	−1.2
In+	3.3(3)	$<10^{-4}$	−0.017	−0.16
Sr	407.873(11)	0.0698(3)	2.279	−53
Yb	240.269(5)	0.0145(15)	−1.27	−25

Source: All values are taken from Ludlow et al. (2015), other than the values for Sr and Yb which were taken from Nicholson et al. (2015) and Sherman et al. (2012), respectively.

atomic species, is known. In Table 9.1, we list the BBR coefficients for commonly used optical clock species and the corresponding BBR shifts at room temperature.

9.3.2.4 Doppler Shifts

The motion of an atom with respect to the laboratory frame will result in a shift in the measured resonance frequencies of its transitions. An atom moving with velocity $\vec{\upsilon}$with respect to a probe laser with wavevector $\vec{k}$, and with a "true" transition frequency ν_o for an atom in its rest frame, will absorb photons from the probe laser at a laser frequency in the lab frame $\nu = \nu_o + \Delta\nu$, where the fractional shift $\Delta\nu/\nu_o$ is given by

$$\frac{\Delta\nu}{\nu_o} \approx \frac{\vec{k}\cdot\vec{\upsilon}}{c} - \frac{|\vec{\upsilon}|^2}{2c^2} + \frac{|\vec{k}\cdot\vec{\upsilon}|^2}{c^2} + O\left(\frac{\upsilon}{c}\right)^3, \tag{9.4}$$

where c is the speed of light (Ludlow et al. 2015).

The first term in equation (9.4) is the first-order Doppler shift. For atoms in free space, the first-order Doppler shift results in significant line broadening of their optical transitions due to their Maxwell–Boltzmann velocity distribution. Even for atoms cooled to a temperature of 1 μK, roughly the minimum achievable with laser cooling, the Doppler broadening of an optical transition is tens of kHz. This is the primary reason atoms in an optical clock must be confined. If an atom is confined tightly enough the internal transitions become decoupled from the atomic motion and Doppler broadening can be eliminated (Ido & Katori 2003). However, once the atoms are confined in a trap, a first-order Doppler shift can still arise if the trapping potential itself is moving with respect to the laser, and in ion clocks this must be carefully monitored. In optical lattice clocks this shift can be avoided by referencing the phase of the clock laser to the mirror used to retro-reflect the lattice (Bloom et al. 2014).

The second and third terms in equation (9.4) correspond to the second-order, or relativistic, Doppler shift, which can be interpreted as the special-relativistic time-dilation of the atom's frequency due to its motion with respect to the lab frame. While this effect is small for free-space atoms cooled to μK temperatures (below the 10^{-20} level for the fractional frequency shift of an optical transition), in ion clocks the large trapping frequencies and ion micromotion can result in large velocities even for an ion cooled into the ground state of the trap, and uncertainty in the second-order Doppler shift can be a dominant systematic effect. In optical lattice clocks, the trapping frequencies are low enough that the second-order Doppler shift is not currently relevant.

In addition to the Doppler shifts given in equation (9.4), unconfined atoms will experience an additional "recoil shift" due to the momentum kick an atom receives when it absorbs a photon, given by $\Delta\nu = \hbar k^2/(4\pi m)$, where m is the mass of the atom and k is the wave number of the laser. This shift can be significant for optical transitions, particularly for light atoms, and the resulting motion can result in further complications. The recoil shift therefore provides additional motivation for confining atoms in a trapping potential. For tightly-confined atoms in the "Lamb-Dicke" regime, the momentum kick from the photons is absorbed by the trap, not the atoms, and the recoil shift is strongly suppressed (Ido & Katori 2003).

9.3.2.5 Gravitational Redshift

The theory of general relativity predicts that a difference in gravitational potential between two clocks will result in a gravitational redshift of the relative clock frequencies. On Earth, this means that if one clock is raised with respect to the other, the clock further from the centre of the Earth will appear to tick faster than the lower clock. Expanding the potential to first order, the predicted fractional frequency shift between two clocks for a small height difference Δh and a local acceleration due to gravity $g \approx 9.8\ \mathrm{m\,s^{-2}}$ is given by

$$\frac{\Delta\nu}{\nu_o} \approx \frac{g\Delta h}{c^2}, \tag{9.5}$$

where c is the speed of light (Ludlow et al. 2015). This corresponds to a fractional frequency shift of 1×10^{-18} at a 1-cm height difference. This effect has now been observed at a height difference of 30 cm using two trapped ion optical clocks (Chou, Hume, Rosenband & Wineland 2010). With clock comparisons now achieving the 10^{-18} level, it has become necessary to measure the relative differences in clock altitude at the mm scale in order to properly account for the gravitational redshift effect, which is pushing the limits of the capabilities of geodetic surveys. While this is problematic for optical clock comparisons, it also presents the opportunity to use optical clocks as a precision tool to measure the Earth's geoid, as discussed in Section 9.4.2. It is worth noting that unlike most other systematics, the gravitational redshift is not a true shift of the clock frequency but rather is a signature of the relative nature of time in a gravitational potential.

9.3.3 Single-Ion Optical Clocks

Ion clocks operate by trapping charged atoms in a harmonic potential at the field minimum of an oscillating RF electric field. Because the ion is extremely well isolated, very tightly confined, and can be cooled to its motional ground state, ion clocks offer narrow clock transition linewidths and excellent control over systematic shifts. Unfortunately, due to the strong Coulomb interaction between ions, at present only one clock ion per harmonic potential is used, limiting the atom number in equation (9.1) to $N = 1$. Nevertheless, single-ion optical clocks with stabilities far surpassing those of microwave atomic clocks and accuracies at the 10^{-18} level have been demonstrated (Huntemann et al. 2016).

9.3.3.1 Ion Clock Operation

Thanks to their charge, ions can be strongly confined using electric fields. Unfortunately, Maxwell's equations forbid the stable trapping of a charged particle in all three dimensions using only dc fields. To circumvent this, in a radio frequency (RF) Paul trap oscillating electric fields are combined with static fields to generate a time-averaged 3D harmonic trapping potential. Figure 9.4 shows some example ion clock Paul trap geometries from the NIST ion clock group in Boulder.

In an ion clock, neutral atoms are produced from a nearby heated source and are then ionized through resonant photoionization. The large trap depths in an RF Paul trap make it possible to then capture single ions in the trap directly, without requiring prior slowing or cooling. Once trapped, a single ion can be held for hours or even months at a time (Ludlow et al. 2015). For clock ions with convenient strong cycling transitions, such as Sr+, Yb+, Ca+, and Hg+, the ion is then laser-cooled using a combination of Doppler cooling and resolved sideband cooling to close to its motional ground state. At this point the ion is ready to be interrogated using one of the clock measurement sequences described in Section 9.3.1. The fluorescent read-out of the clock state population at the end of a clock measurement results in heating of the ion motional state, and the ion must be re-cooled using laser cooling before it can be probed with the clock laser again.

Research into trapped-ion based quantum computation (Leibfried et al. 2003) has led to the development of "quantum logic spectroscopy," which enables the use of ions with

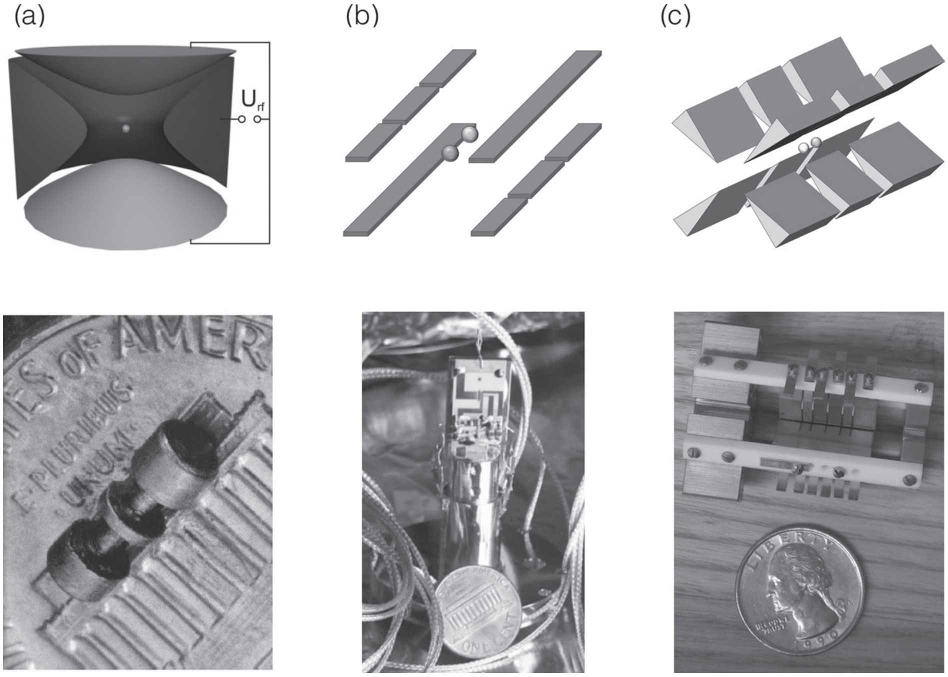


FIGURE 9.4 Examples of RF Paul ion trap geometries. (a) The spherical Paul trap used for the Hg+ frequency standard at NIST. The diagram is from Ludlow et al. (2015) and illustrates the cylindrically symmetric oscillating potential U_{RF} confining the Hg+ ion in red, which is applied between the ring and the end cap electrodes. (b) The linear chip Paul trap used for one of the two Al+ frequency standards at NIST, with the Be+ logic ion in orange and the Al+ clock ion in green. (c) The linear blade Paul trap used for the second Al+ frequency standard at NIST, with the Mg+ logic ion in yellow. All pictures of the traps are taken from the NIST image gallery, and the synergies associated with NIST's position within the US Department of Commerce are also apparent.

desirable clock properties such as small BBR coefficients and narrow clock transition linewidths, but which lack a convenient optical cycling transitions required for laser cooling and electron-shelving-based state read-out (Schmidt et al. 2005). An example is Al+, for which the $1S_0$ to $3P_0$ clock transition has an 8 mHz linewidth and an anomalously low BBR coefficient (see Table 9.1), but for which the strong $1S_0$ to $1P_1$ cycling transition is deep in the UV at 169 nm, and is therefore currently inaccessible. Quantum logic spectroscopy works by trapping a single clock ion such as Al+ with a "logic" ion in the same harmonic-trapping potential. The logic ion is chosen for its efficient cooling transitions and high-fidelity internal state read-out. Thanks to the strong coulomb interaction between the clock ion and the logic ion, Doppler and sideband cooling of the logic ion can be used to sympathetically cool the clock ion. In addition, after the clock ion has been probed with the clock laser, a two-qubit quantum gate can be applied to transfer the internal state of the clock ion into the logic ion to be read out (Schmidt et al. 2005). Figure 9.5 shows the outcome of a recent comparison between two Al+ clocks that make use of quantum logic spectroscopy, demonstrating a single-clock stability of $\sigma_y(\tau) = 2.8 \times 10^{-15}/\sqrt{\tau}$ (Chou, Hume, Koelemeij, Wineland & Rosenband 2010).

9.3.3.2 Ion Clock Systematics

The current generation ion clocks are primarily limited by the systematic effects discussed in Section 9.3.2. Second-order Doppler shifts arising from the micromotion and the harmonic, or "secular," motion of the ion in its tight trapping potential are particularly significant, but are mitigated by the remarkable levels of precision with which the motional states of a single ion can be characterized and controlled. A representative systematic evaluation for a Yb+ ion clock is shown in Table 9.2, published in Huntemann et al. (2016). The dominant sources of systematic uncertainty were the second-order Doppler shifts, the remaining blackbody shift, and the ac Stark shift from the clock laser itself. A total uncertainty of 3.2×10^{-18} was achieved by taking advantage of a second narrow linewidth transition in Yb+ with stronger coupling to magnetic and electric fields in order to evaluate the field-induced shifts of the clock transition, including the blackbody shift.

TABLE 9.2

Evaluated Systematic Shifts for a Yb+ Ion Clock at PTB

Effect	Shift (10^{-18})	Uncertainty (10^{-18})
Second-order Doppler shift	−3.7	2.1
BBR shift	−70.5	1.8
Probe light related shift	0	1.1
Second-order Zeeman shift	−40.4	0.6
Quadratic dc Stark shift	−1.2	0.6
Background-gas collisions	0	0.5
Servo error	0	0.5
Quadrupole shift	0	0.3
Total	−115.8	3.2

Source: All values are taken from Huntemann et al. (2016).

9.3.4 Neutral Atom Optical Lattice Clocks

In an optical lattice clock thousands of neutral atoms are confined in a trapping potential generated by the interference pattern of counter-propagating laser beams. The strong confinement provided by the lattice isolates the narrow internal optical clock transition from the motion of the atoms. Because they are neutrally charged, the atoms interact only weakly (Martin et al. 2013) and can all be probed simultaneously with the same probe laser, increasing the achievable N over ion clocks and thereby offering record-setting stabilities (Schioppo et al. 2017, Campbell et al. 2017). To date, absolute frequency measurements of optical lattice clocks with the alkaline-earth element strontium and the alkaline-earth-like elements ytterbium and mercury have been demonstrated. While the increased number of atoms and the use of an optical lattice results in additional systematic shifts that must be evaluated and controlled, optical lattice clocks have also recently demonstrated accuracies at the 10^{-18} level (Nicholson et al. 2015).

9.3.4.1 Optical Lattice Clock Operation

In a typical OLC, hot atoms are loaded from an oven source into the clock chamber through a "Zeeman slower," so that by the time they reach the chamber the atoms are near rest along the axis of the slower. Once inside the clock chamber,

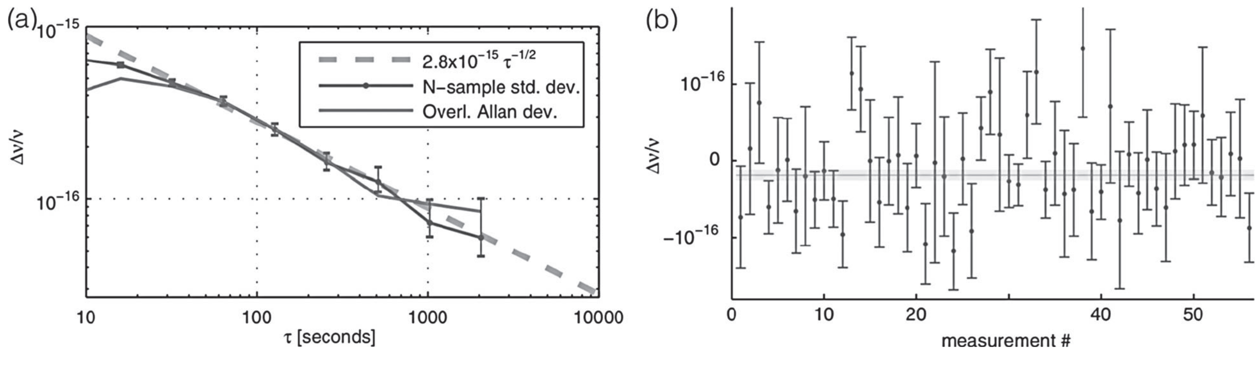


FIGURE 9.5 **Comparison between two quantum logic ion optical clocks.** (a) Fractional frequency uncertainty vs averaging period for a comparison between the two Al+ clocks at NIST. (b) Measurements of the fractional frequency difference between the two clocks Al+ (blue points). Error bars represent the statistical uncertainty. The horizontal line shows the weighted mean of the difference between the two clocks, -1.8×10^{-17}, and the shaded yellow bar the standard deviation of 7×10^{-18}. (Figures from Chou, Hume, Koelemeij, Wineland & Rosenband 2010.)

the atoms are captured and cooled in one or two stages of magneto-optical-traps (MOTs), a common atomic physics technique that combines 3D laser cooling with a conservative potential achieved using magnetic field gradients generated by anti-Helmholtz coils (Metcalf & Van der Straten 2007) (Figure 9.6).

After MOT cooling, the atoms are sufficiently cold and localized to be trapped by the optical dipole force in a one-dimensional optical lattice formed by a laser beam interfering with its counter-propagating retro-reflection. The importance of the strong optical confinement provided by the lattice is illustrated in the data shown in Figure 9.7b. When there is no axial trapping potential, (red diamonds), the clock transition has a Doppler-broadened linewidth due to the finite temperature of the atoms. As the axial trapping potential is increased, the motional sidebands become spectrally resolved, and the clock carrier transition at zero detuning narrows. In a deep lattice (blue squares) the clock is in the side-band-resolved and Lamb-Dicke regimes, and the finite temperature of the atoms has a negligible effect on the clock transition linewidth.

Unfortunately the trapping light also results in significant shifts of the internal structure of the atoms through the ac Stark effect, potentially leading to troublesome shifts of the clock transition that are dependent on both the lattice intensity and frequency. This can be partially circumvented by using a "magic wavelength" trap, where the frequency of the trapping light is carefully chosen such that the light-dependent frequency shifts of the two clock states cancel to first order. Nevertheless, as discussed in Section 9.3.4.2, the residual lattice shifts must be carefully evaluated.

Once the lattice has been ramped up, thousands of atoms at μK temperatures are confined in two-dimensional pancake-like optical potentials with depths of tens of μK (see Figure 9.7a), and they are ready to be probed on the narrow clock transition using the sequences described in Section 9.3.1 without suffering adverse effects from their finite temperature. Due to the shallow trapping depth, in optical lattice clocks the fluorescence measurements of the clock state populations typically heat the atoms out of the lattice, and new atoms must be cooled and loaded into the lattice for each measurement cycle.

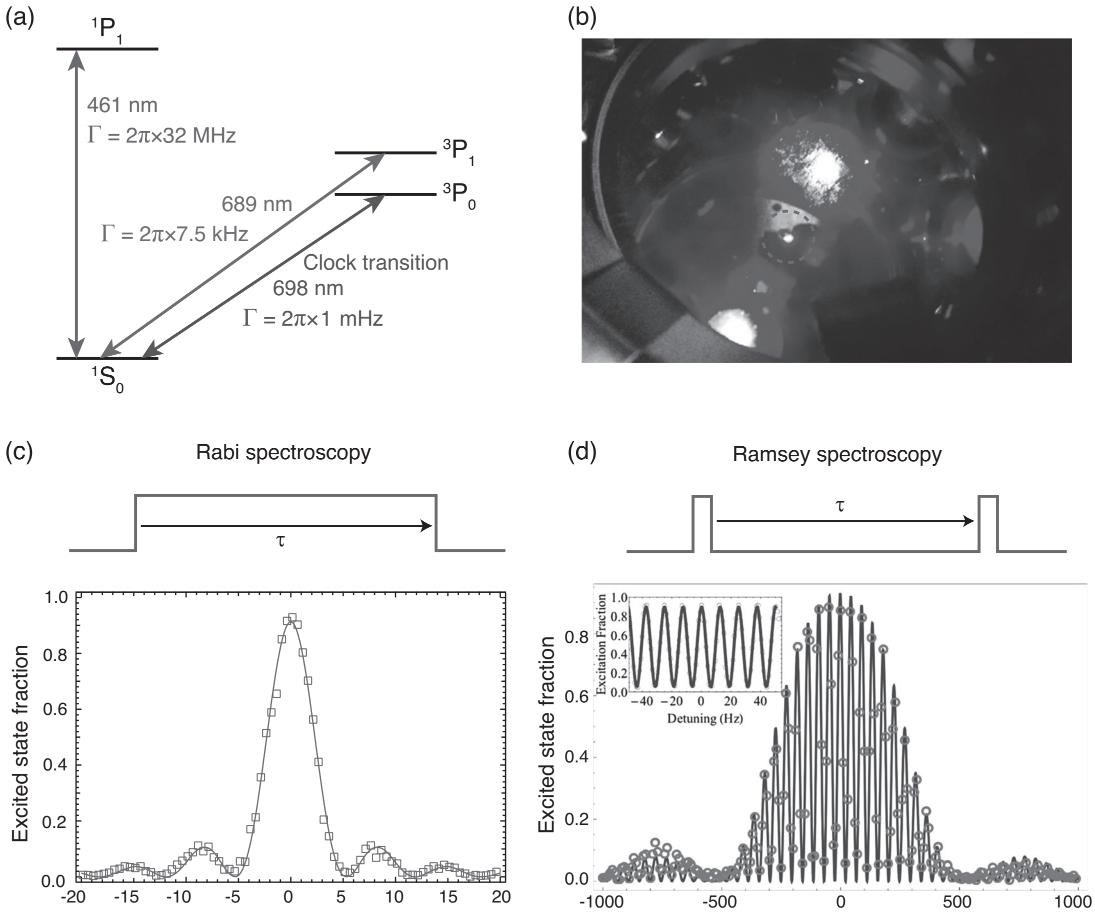


FIGURE 9.6 Standard strontium OLC operation. (a) A simplified level structure of 87Sr with the relevant cooling and clock transitions. (b) A picture of the 87Sr blue MOT in a strontium optical lattice clock in the Ye group at JILA. The 87Sr atoms are visible as a glowing blue ball in the centre of the red dashed circle. (c) A Rabi scan of the clock transition in the deep lattice regime, taken with the same strontium optical lattice clock. Here, the observed linewidth is a result of the Fourier limit for the applied 160 ms π-pulse. (d) The measured excitation fraction versus detuning for Ramsey spectroscopy where detuning is scanned over a larger range than the Rabi frequency. The inset shows the measured excitation fraction versus detuning using Ramsey spectroscopy for a π/2-pulse time of 750 μs and a free evolution time of 80 ms that is scanned ±50 Hz from resonance. (Figure from Refs. Martin 2013 and Bishof 2014.)

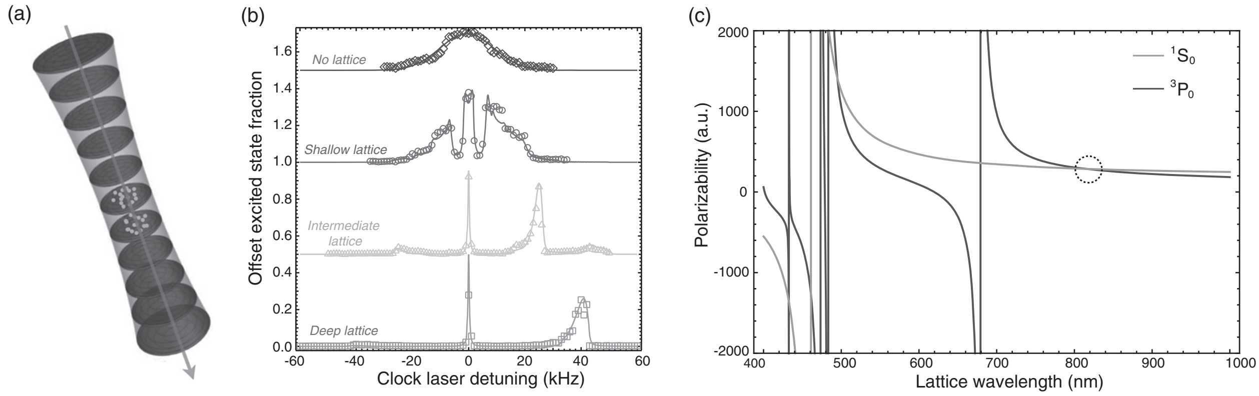


FIGURE 9.7 Strong confinement in an optical lattice. (a) A cartoon of atoms inside a 1D optical lattice, probed with a clock laser along the lattice axis. (b) Clock spectroscopy of a strontium optical lattice clock in the Ye group at JILA, published in Kolkowitz et al. (2017), showing the impact of the depth of the magic wavelength optical lattice on the 698-nm clock transition in a horizontal lattice. The axial motional sidebands of the lattice become well resolved from the narrow clock transition at zero detuning as the depth of the lattice is increased. The pronounced asymmetry between the red and blue sidebands illustrates that the atoms are primarily in the ground band on the lattice. (c) Polarizability curves for the two neutral strontium clocks states ($1S_0$ in blue, $3P_0$ in red). The "magic wavelength" at 813 nm is highlighted in the black dashed circle.

In Figure 9.8, a comparison between two ytterbium optical lattice clocks is shown, demonstrating a single-clock stability of $\sigma_y(\tau) = 3.2\times10^{-16}/\sqrt{\tau}$, and $\sigma_y = 1.6\times10^{-18}$ at 25000 s (Hinkley et al. 2013).

9.3.4.2 Optical Lattice Clock Systematics

A representative systematic evaluation for a strontium optical lattice clock is shown in Table 9.3. In addition to the systematics already discussed in Section 9.3.2, one of the unique systematics for optical lattice clocks is the ac Stark shift from the lattice light (Brown et al. 2017). While the lattice frequency and polarization are carefully selected to operate as close as possible to the "magic wavelength" condition, uncertainty in these parameters can result in a significant systematic shift. In the limit where the ac Stark shift depends only linearly on the lattice trap depth U_o, the full expression for the lattice frequency shift of the clock transition $\Delta\nu_{\text{lattice}}$ as a function of the lattice frequency f is given by

$$\begin{aligned}\Delta\nu_{\text{lattice}} = U_o\Big\{&\Delta\kappa_s(f) + \Delta\kappa_\upsilon(f)m_f\xi\vec{k}\cdot\vec{B}\\ &+[3m_f^2 - F(F+1)]\left(3\left|\vec{\varepsilon}\cdot\vec{B}\right|^2-1\right)\Delta\kappa_t(f)\Big\}\end{aligned}\tag{9.6}$$

where $\vec{k}$ is the lattice propagation vector, $\vec{\varepsilon}$ is the lattice polarization vector, ξ is the lattice polarization ellipticity (0 indicates linear polarization), $\vec{B}$ is the bias magnetic field direction,

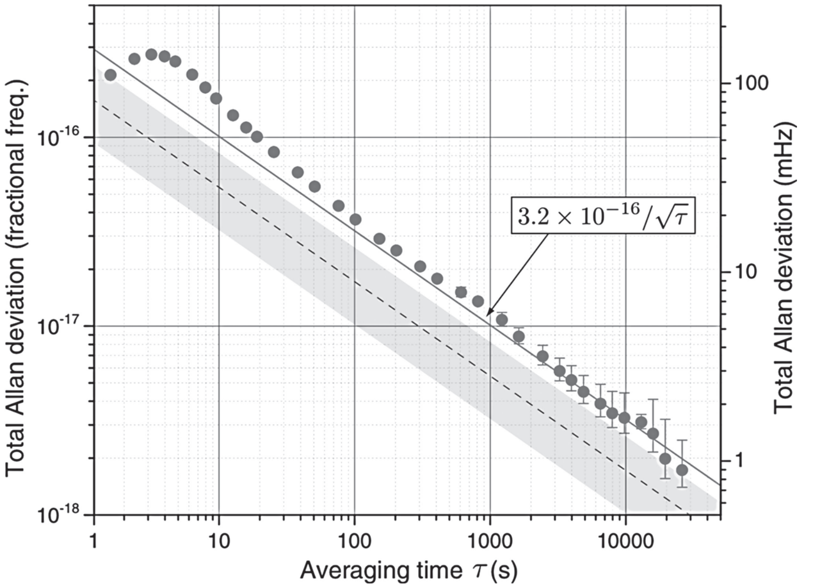


FIGURE 9.8 Comparison between two Yb optical lattice clocks. Extracted single-clock fractional frequency uncertainty and frequency uncertainty vs averaging time for a comparison between the two optical lattice clocks at NIST. The blue dashed line represents the estimated combined instability contributions from the Dick effect; the shaded region denotes uncertainty in these estimates. (This figure is from Hinkley et al. 2013 and is reprinted with permission from AAAS.)

TABLE 9.3

Evaluated Systematic Shifts for a Strontium Optical Lattice Clock at JILA

Effect	Shift (10^{-18})	Uncertainty (10^{-18})
Lattice ac Stark shift	−1.3	1.1
BBR static shift	−4562.1	0.3
BBR dynamic shift	−305.3	1.4
dc Stark shift	0	0.1
Probe ac Stark shift	0	0.0
First-order Zeeman shift	−0.2	0.2
Quadratic Zeeman shift	−51.7	0.3
Density shift	−3.5	0.4
Line-pulling+tunnelling	0	<0.1
Second-order Doppler shift	0	<0.1
Background gas collision shift	0	<0.6
Servo offset shift	−0.5	0.4
AOM phase chirp	0.6	0.4
Total	−4929.0	2.1

Source: All values are taken from Nicholson et al. (2015).

F is the total angular momentum quantum number, m_f is the hyperfine spin state, and $\Delta\kappa_s$, $\Delta\kappa_v$, and $\Delta\kappa_t$ are the differential scalar, vector, and tensor shift coefficients, respectively (Nicholson et al. 2015). The vector Stark shift can be doubly suppressed through a combination of linearly polarized lattice light with $\xi = 0$ and by aligning the lattice perpendicular to the magnetic field such that $\vec{k} \cdot \vec{B} = 0$. The "magic wavelength" is then the particular frequency f' for which the scalar and tensor shifts cancel and $\Delta\nu_{\text{lattice}} = 0$ for the operating conditions of the clock. Using this technique lattice Stark shift uncertainties at the 1.1×10^{-18} level have been demonstrated (Nicholson et al. 2015). At these levels of uncertainty the higher order nonlinear scaling of the ac Stark shift with lattice depth due to the atomic hyperpolarizability must also be accounted for (Brown et al. 2017), and reaching lower levels of uncertainty with optical lattice clocks will likely require new techniques to mitigate the ac Stark shift, as we discuss in Section 9.4.1.1.

Another important systematic shift in optical lattice clocks is the density shift due to collisional interactions between atoms occupying the same lattice site. These interactions not only introduce a significant frequency shift but also result in line-broadening that can limit the coherent interrogation times (Bloom et al. 2014). The loading of the lattice from the MOT results in a Poissonian probability distribution for the number of atoms on each site (Swallows et al. 2012). For each measurement sequence each lattice site therefore has a different density and a different corresponding collisional shift, leading to inhomogeneous broadening of the clock line, and to line pulling and corresponding systematic uncertainty. Until recently, the principle approach to this problem has been to decrease the density using the largest possible lattices and keeping the atom number low, and then to evaluate the density shift by interleaving measurements between high and low densities and performing a self-comparison (Bloom et al. 2014, Le Targat et al. 2013). However, as described in Section 9.4.1, the use of a Fermi-degenerate 3D optical lattice to greatly increase atomic density whilst effectively eliminating the onsite interaction shift has recently been demonstrated (Campbell et al. 2017). With single-site interactions controlled or eliminated, it is likely that the collective dipolar interactions between atoms on nearby lattice sites will now become significant (Bromley et al. 2016). This constant emergence of new, ever more complicated systematics is a common theme in optical clock research, and whilst it may seem vexing, it also offers exciting opportunities for new discoveries and clock applications, as we discuss in the following section.

9.4 Outlook and Future Directions

9.4.1 Next-Generation Clocks

In the prior sections of this chapter we have discussed a number of factors that currently limit the stability and accuracy of optical clocks. Fortunately, despite the fact that optical clocks are already operating at remarkable levels of precision, there are a number of emerging next-generation optical clock technologies that promise to enable further gains. Here we give a brief overview of some of these promising developments.

9.4.1.1 3D Optical Lattice Clocks

As described in Section 9.3.4, the current generation of optical lattice clocks makes use of 1D optical lattices, in which multiple neutral atoms are loaded into each pancake-shaped lattice site. The atoms are therefore in well-defined quantized motional states along the axis of the lattice, but their motion is thermal along the weakly confined transverse dimensions. The resulting collisional interactions between atoms occupying the same lattice sites give rise to an unfortunate trade-off between the atom number N and the coherent interrogation time and magnitude of the density shift, limiting both the stabilities and accuracies of these clocks. However, by adding additional lattice beams, strong confinement of clock atoms in two and even three dimensions has previously been achieved (Swallows et al. 2011, Akatsuka et al. 2010). It was recently demonstrated that the strong confinement provided by a 3D lattice clock results in spectrally well-resolved on-site interactions that eliminate line-broadening and line-pulling from collisional interactions (Campbell et al. 2017). Furthermore, the loading of a 3D optical lattice clock from a Fermi-degenerate gas offers the prospect of millions of atoms deterministically loaded into the lattice with unit filling factor, which each atom in a well-defined quantum state like an ion clock. When used for differential clock comparisons or in concert with new ultra-stable optical cavities, 3D lattice clocks will likely provide order of magnitude gains in clock stability. Furthermore, the introduction of high numerical aperture optical imaging of the lattice offers new possibilities for the characterization and control of systematics, as well as the single-shot measurement of Rabi lines and Ramsey fringes to characterize laser noise spectra (Marti et al. 2018). Of course, the evolution of optical lattice clock geometry is unlikely to stop here. The observation of Raman scattering out of the excited clock state due to the high-intensity 3D lattice beams (Campbell et al. 2017) has motivated new proposals for 3D "accordion lattices" formed by beams at acute angles. These lattices will have large, tunable spacings between the lattice sites, enabling clock operation at low lattice intensities without the detrimental effects of tunnelling (Hutson et al. 2018).

9.4.1.2 Cryogenic Optical Clocks

As described in Section 9.3.2, one of the primary limiting systematic effects in both ion and optical lattice clocks is the shift due to BBR. To reach the 10−18 level of systematic uncertainty, optical clocks have already had to employ atoms with especially small blackbody coefficients (Chou, Hume, Koelemeij, Wineland & Rosenband 2010), the use of the most accurate thermometers available (Nicholson et al. 2015), or the use of an in-vacuum radiation shield (Beloy et al. 2014). However, the scaling of this shift with the temperature of the surrounding environment (T^4 for the static contribution and T^6 for the dynamic contribution,) means that by moving from atoms surrounded by 300 K ambient room temperature to the 4 K cryogenic temperatures achievable in a helium-based cryostat, this systematic can be suppressed by many orders of magnitude. This is difficult to implement because the atoms must be entirely enclosed by cryogenic radiation shields, as even a very small solid-angle opening to surfaces at warmer temperatures can result in large uncertainties (Ushijima et al. 2015). Despite the technical challenges, the first generation of cryogenic ion clocks and optical lattice clocks have already been realized (Ushijima et al. 2015, Rosenband et al. 2008).

9.4.1.3 Superradiant Optical Clocks

The challenges associated with building ever more stable passive optical cavities to reduce clock laser linewidths and mitigate the Dick effect (see Section 9.2.1) motivate the search for other approaches. One of the most promising alternatives is the development of an active optical frequency reference, in analogy to the hydrogen masers that act as active references in the microwave regime and offer excellent short-term stability to complement the long-term stability of microwave atomic clocks. The optical equivalent is a superradiant ultra-narrow linewidth laser, in which neutral atoms are placed inside an optical cavity tuned to resonance with a narrow atomic clock transition (Meiser et al. 2009, Bohnet et al. 2012). Atoms pumped into the excited clock state will then experience superradiant collective emission into the cavity mode, with the coherence of the light determined not by the cavity but by the atoms. The first demonstration of pulsed superradiant emission with an ultra-narrow optical clock transition was recently performed with strontium (Norcia et al. 2018). Efforts to extend these encouraging results to continuous wave (CW) superradiant emission are ongoing, with prospects for use both to enhance the performance of traditional optical clocks, and as a complementary optical clock technology with outstanding short-term stability.

9.4.1.4 Entangled Clocks

While equation (9.1) from Section 9.1.1 represents the quantum projection noise (QPN) limit to the stability of an optical clock with N independent atoms, it is in principle possible to further enhance the clock stability by placing the atoms in non-classical entangled states such as spin-squeezed states and maximally entangled Greenberger-Horne-Zeilinger states (Wineland et al. 1992, Leroux et al. 2010, Bohnet et al. 2014, Kessler et al. 201$\sqrt{4}$). This can push the stability of the clock below the QPN limit given by equation (4.8.1), which scales with $1/\sqrt{N}$, towards the ultimate Heisenberg limit allowed by quantum mechanics (Wineland et al. 1992), which scales with $1/N$. This is particularly attractive in light of the increased atom numbers promised by the 3D lattice clocks discussed in Section 9.4.1.1. However, in order to generate entanglement between the internal states of the atoms, a method for engineering strong, controllable atom–atom interactions is required. Proposed approaches include the use of high-finesse optical cavity to generate strong light-mediated atom–atom interactions (Norcia & Thompson 2016), or the use of Rydberg states to enhance the strength and length-scale of the atom–atom interactions (Gil et al. 2014, Kómár et al. 2016). Once entangled states have been generated they will be particularly sensitive to sources of decoherence including laser noise, and so they may prove to be most useful for clocks with small numbers of atoms such as ion clocks, or for differential clock comparisons in which the laser noise can be eliminated (Marti et al. 2018).

9.4.1.5 Exotic Atomic Clocks

While the optical atomic clocks described in this chapter are the most developed and widely adopted frequency standards, new kinds of exotic clocks are just beginning to emerge as next-generation optical clock candidates. For example, electronic transitions in highly charged ions (Safronova et al. 2014) and the intra-nuclear transition in 229Th (Campbell et al. 2012) have been proposed as promising clock transitions due to their insensitivity to external perturbations and their higher frequencies. With the recent demonstration of the trapping and sympathetic cooling of individual highly charged ions (Schmöger et al. 2015), there is now a zoo of highly charged ions and corresponding transitions to be explored spanning the entire periodic table. Furthermore, the prospects for enhanced sensitivity to variations in fundamental constants have motivated proposals for the use of clock transitions in highly charged ions and weakly bound molecules in searches for new physics (Safronova et al. 2014, Zelevinsky et al. 2008), as discussed in Section 9.4.2.

9.4.2 Emerging Applications of Optical Clocks

9.4.2.1 SI Unit Definitions, the World Clock, Navigation, and Geodesy

The most obvious application of optical clocks is to the definition and measurement of time and frequency. The SI second is currently defined with respect to the hyperfine transition of caesium. Because time can be defined more precisely than other units, in the new SI unit system currently being adopted, almost all other physical units are defined with respect to this definition of the second and fundamental constants. A worldwide collaboration of standard institutes and laboratories operate the microwave atomic clocks that are averaged together to form the International Atomic Time that serves as the world clock. Optical clocks have already demonstrated accuracies surpassing caesium standards, and it seems clear that eventually an optical atomic transition will be adopted as the new primary definition of the second. However, for time being optical clocks are not yet sufficiently robust or widely available to warrant this switch. In the meantime, several optical transitions have been recognized as "secondary representations of the second," and optical clocks will soon contribute to steering the world clock (Ludlow et al. 2015).

Optical clocks will also likely begin to have a technological impact beyond timekeeping, with potential applications to navigation and geodesy. Microwave atomic clocks already play a central role in the Global Navigation Satellite Systems such as the Global Positioning System (GPS), networks of satellites which provide position and navigation information around the world. Although the positioning accuracy of the GPS is not presently limited by the stability of the vapour cell microwave atomic clocks in the satellites, these clocks require regular hourly corrections from the ground control stations in order for the network to remain synchronized and not accumulate significant errors. Adding additional space-based optical clock satellite nodes would improve GPS system integrity because the network would require less frequent ground-to-space based communications for clock correction and could potentially operate autonomously for days or weeks at a time (Ludlow et al. 2015).

As discussed in Section 9.3.2, the sensitivity of optical clock comparisons to changes in the gravitational potential presents a challenge for high-accuracy clock comparisons but also offers an opportunity to harness optical clocks for "relativistic geodesy." Once optical clock comparisons over long distances have reached the point where geodetic surveys are no longer sufficient to measure the changes in gravitational potential between optical clocks at their levels of relative accuracy, the clocks themselves will become the best tools for the job. Portable optical clocks could be compared locally, then moved to two distant locations and compared over a fibre or free-space optical link to determine their relative height, even between continents. Indeed, the first demonstrations of geodesy measurements between clocks connected over fibre links have already taken place (Takano et al. 2016, Lisdat et al. 2016).

9.4.2.2 Searches for Variations of Fundamental Constants, Dark Matter, and Gravitational Waves

As optical clocks continue to improve in accuracy, their sensitivity to signals of interest for fundamental physics research also grows. For example, because the energies of atomic clock transitions depend on fundamental constants such as the electron–proton mass ratio and the fine structure constant, comparisons between clock transitions that scale differently with these constants, such as comparisons between different atomic species, between microwave and optical clocks, or between two transitions in the same species with different character as in the Yb+ ion, can probe whether these fundamental constants are changing with time (Safronova et al. 2017). Long time-scale comparisons over many years have already placed constraints on slow variations (Huntemann et al. 2014), whilst short time-scale comparisons can search for faster coherent oscillations that might be induced by certain ultra-light dark-matter candidates (Arvanitaki et al. 2015).

Another promising approach for the future is to separate two identical space-based optical clocks over a large baseline and to compare their relative frequencies over an optical link. This type of detector would be sensitive to variations in fundamental constants as a function of position due to the motion of the satellites through a background dark-matter field (Arvanitaki et al. 2016), or to changes in the light travel time between the two clocks due to passing low-frequency gravitational waves (Kolkowitz et al. 2016). The sensitivity of this class of detector can be further enhanced using differential comparisons between the clocks that cancel sensitivity to laser noise, as well as measurement sequences that lock-in to specific frequencies (Kolkowitz et al. 2016).

9.4.2.3 Many-Body Quantum Physics

While the density shift in optical lattice clocks is a troublesome systematic that is typically minimized by operating at low densities, the many-body quantum physics underlying it has proven to be complex and rich (Martin et al. 2013). In addition, the inherent insensitivity to external perturbations and high degree of control demanded by high-accuracy optical clocks is ideal for the characterization of complex interactions and dynamics. As a result, precision clock spectroscopy is emerging as a powerful new tool for the study of exotic quantum systems. A wide range of interesting physics has already been observed, including SU(N) symmetry (Zhang et al. 2014), spin-orbit coupling (Kolkowitz et al. 2017, Livi et al. 2016), and interaction-induced spin-rephasing (Bromley et al. 2018, Norcia et al. 2017). The recent demonstration of a Fermi-degenerate 3D lattice clock (see Section 9.4.1.1) is certain to add access to additional phenomena, such as long-range dipolar interactions and many-body phase transitions.

REFERENCES

Akatsuka, T., Takamoto, M. & Katori, H. (2010), 'Three-dimensional optical lattice clock with bosonic Sr 88 atoms', *Physical Review A* **81**(2), 023402.

Arvanitaki, A., Graham, P. W., Hogan, J. M., Rajendran, S. & Van Tilburg, K. (2016), 'Search for light scalar dark matter with atomic gravitational wave detectors', *arXiv preprint arXiv:1606.04541.*

Arvanitaki, A., Huang, J. & Van Tilburg, K. (2015), 'Searching for dilaton dark matter with atomic clocks', *Physical Review D* **91**(1), 015015.

Beloy, K., Hinkley, N., Phillips, N. B., Sherman, J. A., Schioppo, M., Lehman, J., Feldman, A., Hanssen, L. M., Oates, C. W. & Ludlow, A. D. (2014), 'Atomic clock with 1×10^{-18} room-temperature blackbody Stark uncertainty', *Physical Review Letters* **113**(26), 260801.

Beloy, K., Zhang, X., McGrew, W., Hinkley, N., Yoon, T., Nicolodi, D., Fasano, R., Schäffer, S., Brown, R. & Ludlow, A. (2018), 'Faraday-shielded dc Stark-shift-free optical lattice clock', *Physical Review Letters* **120**(18), 183201.

Bishof, M. (2014), Understanding atomic interactions in an optical lattice clock and using them to study many-body physics, *PhD thesis*, University of Colorado at Boulder.

Bishof, M., Zhang, X., Martin, M. J. & Ye, J. (2013), 'Optical spectrum analyzer with quantum- limited noise floor', *Physical Review Letters* **111**(9), 093604.

Bloom, B. J., Nicholson, T. L., Williams, J. R., Campbell, S. L., Bishof, M., Zhang, X., Zhang, W., Bromley, S. L. & Ye, J. (2014), 'An optical lattice clock with accuracy and stability at the 10^{-18} level', *Nature* **506**(7486), 71–75.

Bohnet, J. G., Chen, Z., Weiner, J. M., Meiser, D., Holland, M. J. & Thompson, J. K. (2012), 'A steady-state superradiant laser with less than one intracavity photon', *Nature* **484**(7392), 78.

Bohnet, J. G., Cox, K. C., Norcia, M. A., Weiner, J. M., Chen, Z. & Thompson, J. K. (2014), 'Reduced spin measurement backaction for a phase sensitivity ten times beyond the standard quantum limit', *Nature Photonics* **8**(9), 731–736.

Bromley, S., Kolkowitz, S., Bothwell, T., Kedar, D., Safavi-Naini, A., Wall, M., Salomon, C., Rey, A. & Ye, J. (2018), 'Dynamics of interacting fermions under spin–orbit coupling in an optical lattice clock', *Nature Physics* 14(4), 399–404.

Bromley, S. L., Zhu, B., Bishof, M., Zhang, X., Bothwell, T., Schachenmayer, J., Nicholson, T. L., Kaiser, R., Yelin, S. F., Lukin, M. D., Rey, A. M. (2016), 'Collective atomic scattering and motional effects in a dense coherent medium', *Nature communications* **7**, 11039.

Brown, R., Phillips, N., Beloy, K., McGrew, W., Schioppo, M., Fasano, R., Milani, G., Zhang, X., Hinkley, N., Leopardi, H., Yoon, T. H. (2017), 'Hyperpolarizability and operational magic wavelength in an optical lattice clock', *arXiv preprint arXiv:1708.08829.*

Campbell, C. J., Radnaev, A. G., Kuzmich, A., Dzuba, V. A., Flambaum, V. V. & Derevianko, A. (2012), 'Single-ion nuclear clock for metrology at the 19th decimal place', *Physical Review Letters* **108**(12), 120802.

Campbell, G. K., Ludlow, A. D., Blatt, S., Thomsen, J. W., Martin, M. J., de Miranda, M. H., Zelevinsky, T., Boyd, M. M., Ye, J., Diddams, S. A., Heavner, T. P. (2008), 'The absolute frequency of the ^{87}Sr optical clock transition', *Metrologia* **45**(5), 539.

Campbell, S. L., Hutson, R., Marti, G., Goban, A., Oppong, N. D., McNally, R., Sonderhouse, L., Robinson, J., Zhang, W., Bloom, B., Ye, J. (2017), 'A Fermi-degenerate three-dimensional optical lattice clock', *Science* **358**(6359), 90–94.

Chou, C., Hume, D., Koelemeij, J., Wineland, D. & Rosenband, T. (2010), 'Frequency comparison of two high-accuracy Al$^+$ optical clocks', *Physical Review Letters* **104**, 070802.

Chou, C., Hume, D., Rosenband, T. & Wineland, D. (2010), 'Optical clocks and relativity', *Science* **329**(5999), 1630–1633.

Christensen, B. T., Henriksen, M. R., Schäffer, S. A., Westergaard, P. G., Tieri, D., Ye, J., Holland, M. J. & Thomsen, J. W. (2015), 'Nonlinear spectroscopy of Sr atoms in an optical cavity for laser stabilization', *Physical Review A* **92**(5), 053820.

Cole, G. D., Zhang, W., Martin, M. J., Ye, J. & Aspelmeyer, M. (2013), 'Tenfold reduction of Brownian noise in high-reflectivity optical coatings', *Nature Photonics* **7**(8), 644.

Cook, S., Rosenband, T. & Leibrandt, D. R. (2015), 'Laser-frequency stabilization based on steady- state spectral-hole burning in Eu^{3+}:Y$_2$SiO$_5$', *Physical Review Letters* **114**(25), 253902.

Cundiff, S. T. & Ye, J. (2003), 'Colloquium: Femtosecond optical frequency combs', *Reviews of Modern Physics* **75**(1), 325.

Diddams, S. A., Bergquist, J. C., Jefferts, S. R. & Oates, C. W. (2004), 'Standards of time and frequency at the outset of the 21st century', *Science* **306**(5700), 1318–1324.

Drever, R., Hall, J. L., Kowalski, F., Hough, J., Ford, G., Munley, A. & Ward, H. (1983), 'Laser phase and frequency stabilization using an optical resonator', *Applied Physics B* **31**(2), 97–105.

Gil, L., Mukherjee, R., Bridge, E., Jones, M. & Pohl, T. (2014), 'Spin squeezing in a Rydberg lattice clock', *Physical Review Letters* **112**(10), 103601.

Hinkley, N., Sherman, J. A., Phillips, N. B., Schioppo, M., Lemke, N. D., Beloy, K., Pizzocaro, M., Oates, C. W. & Ludlow, A. D. (2013), 'An atomic clock with 10^{-18} instability', *Science* **341**(6151), 1215–1218.

Hoyt, C., Barber, Z., Oates, C. W., Fortier, T., Diddams, S. & Hollberg, L. (2005), 'Observation and absolute frequency measurements of the $^1S_0$–$^3P_0$ optical clock transition in neutral ytterbium', *Physical Review Letters* **95**(8), 083003.

Huntemann, N., Lipphardt, B., Tamm, C., Gerginov, V., Weyers, S. & Peik, E. (2014), 'Improved limit on a temporal variation of m_p/m_e from comparisons of Yb$^+$ and Cs atomic clocks', *Physical Review Letters* **113**(21), 210802.

Huntemann, N., Sanner, C., Lipphardt, B., Tamm, C. & Peik, E. (2016), 'Single-ion atomic clock with 3×10^{-18} systematic uncertainty', *Physical Review Letters* **116**(6), 063001.

Hutson, R. et al. (2018), *In preparation*.

Ido, T. & Katori, H. (2003), 'Recoil-free spectroscopy of neutral Sr atoms in the Lamb-Dicke regime', *Physical Review Letters* **91**(5), 053001.

Jiang, Y., Ludlow, A., Lemke, N. D., Fox, R. W., Sherman, J. A., Ma, L.-S. & Oates, C. W. (2011), 'Making optical atomic clocks more stable with 10–16-level laser stabilization', *Nature Photonics* **5**(3), 158.

Kessler, E., M., Komar, P., Bishof, M., Jiang, L., Sorensen, A. S., Ye, J. & Lukin, M. D. (2014), 'Heisenberg-limited atom clocks based on entangled qubits', *Physical Review Letters* **112**, 190403.

Kessler, T., Hagemann, C., Grebing, C., Legero, T., Sterr, U., Riehle, F., Martin, M. J., Chen, L. & Ye, J. (2012), 'A sub-40-mHz-linewidth laser based on a silicon single-crystal optical cavity', *Nature Photonics* **6**(10), 687–692.

Kolkowitz, S., Bromley, S., Bothwell, T., Wall, M., Marti, G. E., Koller, A. P., Zhang, X., Rey, A. M. & Ye, J. (2017), 'Spin-orbit-coupled fermions in an optical lattice clock', *Nature* **542**(20811), 66–70.

Kolkowitz, S., Pikovski, I., Langellier, N., Lukin, M. D., Walsworth, R. L. & Ye, J. (2016), 'Gravi- tational wave detection with optical lattice atomic clocks', *Physical Review D* **94**(12), 124043.

Kómár, P., Topcu, T., Kessler, E., Derevianko, A., Vuletić, V., Ye, J. & Lukin, M. (2016), 'Quantum network of atom clocks: A possible implementation with neutral atoms', *Physical Review Letters* **117**(6), 060506.

Le Targat, R., Lorini, L., Le Coq, Y., Zawada, M., Guéna, J., Abgrall, M., Gurov, M., Rosenbusch, P., Rovera, D. G., Nagórny, B., Gartman, R., Westergaard, P. G., Tobar, M. E., Lours, M., Santarelli, G., Clairon, A., Bize, S., Laurent, P., Lemonde, P. & Lodewyck, J. (2013), 'Experimental realization of an optical second with strontium lattice clocks', *Nature Communications* **4**, 2109.

Leibfried, D., Blatt, R., Monroe, C. & Wineland, D. (2003), 'Quantum dynamics of single trapped ions', *Reviews of Modern Physics* **75**(1), 281.

Leroux, I. D., Schleier-Smith, M. H. & Vuletić, V. (2010), 'Implementation of cavity squeezing of a collective atomic spin', *Physical Review Letters* **104**, 073602.

Lisdat, C., Grosche, G., Quintin, N., Shi, C., Raupach, S., Grebing, C., Nicolodi, D., Stefani, F., Al-Masoudi, A., Dörscher, S., Häfner, S. (2016), 'A clock network for geodesy and fundamental science', *Nature communications* **7**, 12443.

Livi, L., Cappellini, G., Diem, M., Franchi, L., Clivati, C., Frittelli, M., Levi, F., Calonico, D., Catani, J., Inguscio, M., Fallani, L. (2016), 'Synthetic dimensions and spin-orbit coupling with an optical clock transition', *Physical Review Letters* **117**(22), 220401.

Ludlow, A. D., Boyd, M. M., Ye, J., Peik, E. & Schmidt, P. O. (2015), 'Optical atomic clocks', *Reviews of Modern Physics* **87**(2), 637.

Ludlow, A. D., Huang, X., Notcutt, M., Zanon-Willette, T., Foreman, S., Boyd, M., Blatt, S. & Ye, J. (2007), 'Compact, thermal-noise-limited optical cavity for diode laser stabilization at 1×10^{-15}', *Optics letters* **32**(6), 641–643.

Margolis, H., Huang, G., Barwood, G., Lea, S., Klein, H., Rowley, W., Gill, P. & Windeler, R. (2003), 'Absolute frequency measurement of the 674-nm ^{88}Sr^{+} clock transition using a femtosecond optical frequency comb', *Physical Review A* **67**(3), 032501.

Marti, G. E., Hutson, R. B., Goban, A., Campbell, S. L., Poli, N. & Ye, J. (2018), 'Imaging optical frequencies with 100 μ Hz precision and 1.1 μ m resolution', *Physical Review Letters* **120**(10), 103201.

Martin, M. J. (2013), Quantum metrology and many-body physics: pushing the frontier of the optical lattice clock, *PhD thesis*, University of Colorado at Boulder.

Martin, M. J., Bishof, M., Swallows, M. D., Zhang, X., Benko, C., von Stecher, J., Gorshkov, A. V., Rey, A. M. & Ye, J. (2013), 'A quantum many-body spin system in an optical lattice clock', *Science* **341**(6146), 632–636.

Matei, D. G., Legero, T., Häfner, S., Grebing, C., Weyrich, R., Zhang, W., Sonderhouse, L., Robinson, J. M., Ye, J., Riehle, F. & Sterr, U. (2017), '1.5 μm lasers with sub-10 mHz linewidth', *Physical Review Letters* **118**, 263202.

Meiser, D., Ye, J., Carlson, D. & Holland, M. (2009), 'Prospects for a millihertz-linewidth laser', Physical Review Letters **102**(16), 163601.

Metcalf, H. J. & Van der Straten, P. (2007), *Laser Cooling and Trapping of Neutral Atoms*, Wiley Online Library.

Millo, J., Magalhaes, D., Mandache, C., Le Coq, Y., English, E., Westergaard, P., Lodewyck, J., Bize, S., Lemonde, P. & Santarelli, G. (2009), 'Ultrastable lasers based on vibration insensitive cavities', *Physical Review A* **79**(5), 053829.

Nagourney, W., Sandberg, J. & Dehmelt, H. (1986), 'Shelved optical electron amplifier: Observation of quantum jumps', *Physical Review Letters* **56**(26), 2797.

Nazarova, T., Riehle, F. & Sterr, U. (2006), 'Vibration-insensitive reference cavity for an ultra- narrow-linewidth laser', *Applied Physics B* **83**(4), 531.

Nemitz, N., Ohkubo, T., Takamoto, M., Ushijima, I., Das, M., Ohmae, N. & Katori, H. (2016), 'Frequency ratio of Yb and Sr clocks with 5×10^{-17} uncertainty at 150 seconds averaging time', *Nature Photonics* **10**(4), 258.

Nicholson, T., Campbell, S., Hutson, R., Marti, G., Bloom, B., McNally, R., Zhang, W., Barrett, M., Safronova, M., Strouse, G., Tew, W. L. (2015), 'Systematic evaluation of an atomic clock at 2×10^{-18} total uncertainty', *Nature Communications* **6**, 1–8.

Norcia, M. A., Cline, J. R., Muniz, J. A., Robinson, J. M., Hutson, R. B., Goban, A., Marti, G. E., Ye, J. & Thompson, J. K. (2018), 'Frequency measurements of superradiance from the strontium clock transition', *Physical Review X* **8**(2), 021036.

Norcia, M. A., Lewis-Swan, R. J., Cline, J. R., Zhu, B., Rey, A. M. & Thompson, J. K. (2017), 'Cavity mediated collective spin exchange interactions in a strontium superradiant laser', *arXiv preprint arXiv:1711.03673*.

Norcia, M. A. & Thompson, J. K. (2016), 'Strong coupling on a forbidden transition in strontium and nondestructive atom counting', *Physical Review A* **93**(2), 023804.

Poli, N., Oates, C., Gill, P. & Tino, G. (2014), 'Optical atomic clocks', *arXiv preprint arXiv:1401.2378*.

Rosenband, T., Hume, D. B., Schmidt, P. O., Chou, C. W., Brusch, A., Lorini, L., Oskay, W. H., Drullinger, R. E., Fortier, T. M., Stalnaker, J. E., & Diddams, S. A. (2008), 'Frequency ratio of Al^{+} and Hg^{+} single-ion optical clocks; metrology at the 17th decimal place', *Science* **319**(5871), 1808–1812.

Safronova, M., Budker, D., DeMille, D., Kimball, D. F. J., Derevianko, A. & Clark, C. W. (2017), 'Search for new physics with atoms and molecules', *arXiv preprint arXiv:1710.01833*.

Safronova, M., Dzuba, V., Flambaum, V., Safronova, U., Porsev, S. & Kozlov, M. (2014), 'Highly charged ions for atomic clocks, quantum information, and search for α variation', *Physical Review Letters* **113**(3), 030801.

Santarelli, G., Audoin, C., Makdissi, A., Laurent, P., Dick, G. J. & Clairon, A. (1998), 'Frequency stability degradation of an oscillator slaved to a periodically interrogated atomic resonator', *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **45**(4), 887–894.

Schioppo, M., Brown, R. C., McGrew, W. F., Hinkley, N., Fasano, R. J., Beloy, K., Yoon, T. H., Milani, G., Nicolodi, D., Sherman, J. A., Phillips, N. B., Oates, C. W. & Ludlow, A. D. (2017), 'Ultrastable optical clock with two cold-atom ensembles', *Nature Photonics* **11**(1), 48–52.

Schmidt, P. O., Rosenband, T., Langer, C., Itano, W. M., Bergquist, J. C. & Wineland, D. J. (2005), 'Spectroscopy using quantum logic', *Science* **309**(5735), 749–752.

Schmöger, L., Versolato, O., Schwarz, M., Kohnen, M., Windberger, A., Piest, B., Feuchtenbeiner, S., Pedregosa-Gutierrez, J., Leopold, T., Micke, P., & Hansen, A. K. (2015), 'Coulomb crystallization of highly charged ions', *Science* **347**(6227), 1233–1236.

Sherman, J. A., Lemke, N. D., Hinkley, N., Pizzocaro, M., Fox, R. W., Ludlow, A. D. & Oates, C. W. (2012), 'High-accuracy measurement of atomic polarizability in an optical lattice clock', *Physical Review Letters* **108**(15), 153002.

Stenger, J., Tamm, C., Haverkamp, N., Weyers, S. & Telle, H. R. (2001), 'Absolute frequency measurement of the 435.5-nm ^{171}Yb^{+}-clock transition with a Kerr-lens mode-locked femtosecond laser', *Optics letters* **26**(20), 1589–1591.

Swallows, M. D., Bishof, M., Lin, Y., Blatt, S., Martin, M. J., Rey, A. M. & Ye, J. (2011), 'Suppression of collisional shifts in a strongly interacting lattice clock', *Science* **331**(6020), 1043–1046.

Swallows, M., Martin, M., Bishof, M., Benko, C., Lin, Y., Blatt, S., Rey, A. & Ye, J. (2012), 'Operating a ^{87}Sr optical lattice clock with high precision and at high density', *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **3**(59), 416–425.

Takano, T., Takamoto, M., Ushijima, I., Ohmae, N., Akatsuka, T., Yamaguchi, A., Kuroishi, Y., Munekane, H., Miyahara, B. & Katori, H. (2016), 'Geopotential measurements with synchronously linked optical lattice clocks', *Nature Photonics* 10(10), 662–666.

Udem, T., Diddams, S. A., Vogel, K. R., Oates, C., Curtis, E., Lee, W., Itano, W., Drullinger, R., Bergquist, J. & Hollberg, L. (2001), 'Absolute frequency measurements of the Hg^{+} and Ca optical clock transitions with a femtosecond laser', *Physical Review Letters* **86**(22), 4996.

Ushijima, I., Takamoto, M., Das, M., Ohkubo, T. & Katori, H. (2015), 'Cryogenic optical lattice clocks', *Nature Photonics* **9**(3), 185.

Webster, S., Oxborrow, M., Pugla, S., Millo, J. & Gill, P. (2008), 'Thermal-noise-limited optical cavity', *Physical Review A* **77**(3), 033847.

Westergaard, P. G., Lodewyck, J. & Lemonde, P. (2010), 'Minimizing the Dick effect in an optical lattice clock', *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **57**(3), 623–628.

Wineland, D. J., Bollinger, J. J., Itano, W. M., Moore, F. L. & Heinzen, D. J. (1992), 'Spin squeezing and reduced quantum noise in spectroscopy', *Physical Review A* **46**, R6797–R6800.

Yamanaka, K., Ohmae, N., Ushijima, I., Takamoto, M. & Katori, H. (2015), 'Frequency ratio of ^{199}Hg and ^{87}Sr optical lattice clocks beyond the SI limit', *Physical Review Letters* **114**(23), 230801.

Zelevinsky, T., Kotochigova, S. & Ye, J. (2008), 'Precision test of mass-ratio variations with lattice-confined ultracold molecules', *Physical Review Letters* **100**(4), 043201.

Zhang, W., Robinson, J., Sonderhouse, L., Oelker, E., Benko, C., Hall, J., Legero, T., Matei, D., Riehle, F., Sterr, U., & Ye, J. (2017), 'Ultrastable silicon cavity in a continuously operating closed-cycle cryostat at 4 k', *Physical review letters* **119**(24), 243601.

Zhang, X., Bishof, M., Bromley, S. L., Kraus, C. V., Safronova, M. S., Zoller, P., Rey, A. M. & Ye, J. (2014), 'Spectroscopic observation of SU(N)-symmetric interactions in Sr orbital magnetism', *Science* **345**(6203), 1467–1473.

FURTHER READING

Interested readers looking for a more comprehensive and detailed discussion of optical clocks are encouraged to read these two excellent recent review articles:

Ludlow, D., Boyd, M. M., Ye, J., Peik, E., & Schmidt, P. O. (2015), 'Optical atomic clocks,' *Reviews of Modern Physics* 87(2), 637.

Poli, N., Oates, C., Gill, P., & Tino, G. (2014), 'Optical atomic clocks,' *arXiv preprint arXiv:1401.2378.*

10

Optical Atomic Clock and Laser Applications to Length and Time Metrology

Y. Jiang, Y. Huang, and K. Gao

CONTENTS

10.1 Introduction 157
10.2 Outline of Optical Atomic Clocks 158
10.3 Optical Clocks Based on Single Trapped Ion 159
    10.3.1 Ion Trapping 159
    10.3.2 Ion Optical Clocks Referenced to Quadrupole Transitions 160
    10.3.3 Ion Optical Clocks Referenced to the Transitions Other Than the Quadrupole Transitions 161
    10.3.4 Evaluation of Systematic Shifts 162
10.4 Optical Clocks Based on Neutral Atoms 162
    10.4.1 Preparation of Neutral Atoms for Interrogation 163
    10.4.2 Spectroscopy and Laser Locking 164
    10.4.3 Systematic Evaluations 164
10.5 Absolute Frequency Measurement and Optical Frequency Synthesis Using Femtosecond Combs 165
    10.5.1 Absolute Optical Frequency Measurement 166
    10.5.2 Optical Frequency Ratio Measurement 166
    10.5.3 Optical Frequency Synthesis 167
10.6 Internationally Recommended Optical Frequency Standards and Optical Clocks 167
References 168

10.1 Introduction

In the mid-20th century, atoms were realized to be good candidates for time or frequency standards since transition energies or frequencies between two energy levels of some unperturbed atoms are more stable than those defined by the rotation of the earth or artefact-based clocks such as quartz crystal clocks. This motivated extensive research on realization and improvement of such kind of quantum-based time or frequency standards. In 1967, the 13th CGPM (General Conference of Weights and Measures) decided to replace the definition of the second by “the duration of 9 192 631 770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the caesium 133 atom”. Early caesium standards using thermal Cs beams had accuracies of 10^{-12}, whereas Cs fountain clocks at present achieve fractional frequency instabilities of 10^{-15} in 1 day and the lowest frequency inaccuracies of 2×10^{-16} [1], attributing to the development of laser-cooling technique.

The frequency instability of an atomic clock limited by quantum projection noise is given by

$$\sigma(\tau) \approx \frac{\Delta\nu}{\nu}\sqrt{\frac{T_{\text{int}}}{N\tau}} \tag{10.1}$$

where ν is the frequency of the atomic transition, $\Delta\nu$ is the linewidth at the interrogation time of T_{int}, N is the number of absorbers interrogated, and τ is the averaging time. For Cs fountain clocks, $\nu \sim 9.2\times10^9$ Hz, $N \sim 10^6$, $\Delta\nu \sim 1$ Hz, and $T_{\text{int}} \sim 1$ s. Therefore, their frequency instabilities are nearly $10^{-13}/\sqrt{\tau}$. Since optical frequencies (10^{14} Hz) are four orders of magnitude higher than microwave frequencies, atomic clocks based on optical transitions, referred to optical atomic clocks, are more stable than those based on microwave transitions. Motivated by this great advantage, optical atomic clocks have been developed all around the world since 2000.

By choosing extremely weak transitions of atoms, the natural linewidths of the atomic transitions are usually below 1 Hz and even at the millihertz level. Besides a narrow natural linewidth, the coherence time of both the interrogation laser and atoms should be larger than 1 s for resolving a spectral linewidth $\Delta\nu$ below 1 Hz. There are two types of optical atomic clocks. One uses a single ion trapped in a field-free region to obtain a natural-linewidth-resolved atomic transition and an accurate transition frequency insensitive to external perturbations as well. For optical clocks based on a single ion, $N = 1$, and its frequency instability is roughly $10^{-15}/\sqrt{\tau}$. The other employs neutral atoms trapped in optical lattice sites, where N can easily exceed 10^4. Optical lattice clocks have lower quantum-projection-noise-limited frequency instabilities of $10^{-17}/\sqrt{\tau}$.

Nowadays, the frequency instabilities and systematic uncertainties of optical atomic clocks have reached 10^{-18} [2,3]. They have surpassed those of the best caesium fountain clocks who reproduce the current SI unit of time by two orders of magnitude. Eventually the definition of the second will be based on optical atomic clocks, but even now, some of them have been the secondary representations of the second [4]. With such precise and accurate frequency standards, timekeeping and length measurement will enter a new stage. Since the speed of light is fixed as a constant, the metre is defined as "the length of the path travelled by light in vacuum during a time interval of 1/299 792 458 of a second". This definition was adopted in 1983 by the 17th CGPM. Based on this definition, Global Position System (GPS) and Global Navigation Satellite Systems (GNSS) will benefit from high-performance clocks. Meanwhile, the metre can also be realized by means of the wavelength λ in vacuum of a plane electromagnetic wave with the measured frequency f, $\lambda = c/f$, or by means of one of the radiations from the radiations recommended by the International Committee for Weights and Measures (CIPM).

10.2 Outline of Optical Atomic Clocks

An optical atomic clock is based on optical waves with high frequency stability and accuracy. Although lasers can emit optical waves with less frequency noise compared to conventional light sources, free-running laser sources in and of themselves are still too noisy for many clock applications. To reduce the laser frequency noise, lasers are usually frequency-stabilized to an ultra-stable reference cavity (see Figure 10.1). The laser frequency stability is mainly determined by the length stability of the reference cavity, $\delta\nu/\nu \approx -\delta L/L$, where L is the cavity length, δL is the fluctuation of cavity length, and $\delta\nu$ is the laser frequency fluctuation. Any perturbation to the length fluctuation of the reference cavity is minimized, including the vibration, temperature fluctuation, and acoustic noise. The current generation of cavity-stabilized ultra-stable lasers have achieved a frequency instability of 4×10^{-17} at 1 second and a laser linewidth of 10 mHz at 1.5 µm [5].

The ultra-stable laser, referred to as a local oscillator in a clock, still drifts over time due to environmentally induced length fluctuations of the reference cavity. Such drifts can be suppressed by forcing the laser frequency to stay resonant with an atomic transition via feedback systems. The atomic transition could be based on a single ion like Yb^+, Al^+, Sr^+, Ca^+, Hg^+, or it could be based on neutral atoms like Sr, Yb, Hg, Cd. The atomic transition is usually called clock transition, which has a narrow natural linewidth, in order to provide an extremely high-resolution frequency discrimination signal. Moreover, the neutral atoms or a single ion are well confined to a spatial region to reduce the first-order Doppler shift and resolve a natural transition linewidth. They are also isolated from any perturbations to achieve an extreme accurate transition frequency. While locking to reference cavities can provide high stability for short times, second-stage stabilization to a clock transition can yield high frequency stability on longer time scales and high frequency accuracy.

After locked to atomic resonance, the frequency of the local oscillator is stable and accurate enough to serve as an optical frequency standard. However, the oscillation of optical waves is too fast to be counted for any electronic device. Meanwhile as an optical frequency standard, their frequency values must be measured relative to the caesium primary frequency standard at about 9.2 GHz, as this defines the SI second [6]. The large frequency gap between optical wave and microwave can be bridged by an optical frequency comb [7,8]. It serves as a gear in a clock to convert from one frequency to another. It has been measured that optical frequency combs can convert optical frequencies with uncertainties at the 10^{-21} level [9,10],

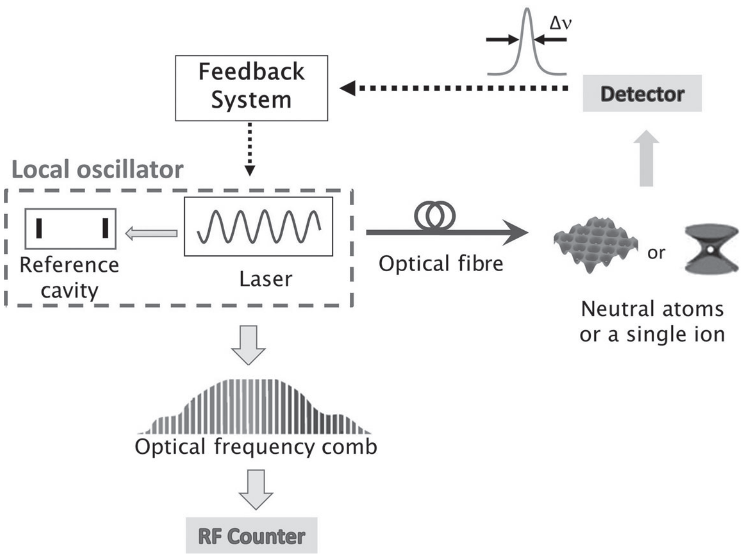


FIGURE 10.1 The schematic diagram of an optical atomic clock.

three orders of magnitude lower than those of the best optical atomic clocks. Therefore, the noise in optical frequency conversion using optical frequency combs can be negligible.

When the different components of an optical atomic clock are significantly separated spatially, they can be connected by optical fibre. Fibres can also be used to transfer the signal of optical atomic clocks to other users or to link two spatially separated labs. However, introduction of optical fibre into the laser path can cause some problems: the optical phase of laser light transferred through the optical fibre is extremely sensitive to environmental perturbation like pressure, temperature, acoustic noise, etc. As a result, the light suffers from random phase modulation, and its spectral purity and accuracy after transfer are degraded. To maintain the phase coherence of the laser light, active fibre phase noise cancellation based on a Michelson interferometer is required [11]. Nowadays, this technique helps to link optical clocks within continents for geopotential measurements and clock comparisons [12].

10.3 Optical Clocks Based on Single Trapped Ion

With the development of ion traps, ultra-narrow linewidth lasers and optical frequency combs, optical clocks based on single trapped ions as a new type of clocks have been realized in many laboratories. Among many possible ion candidates, which are suitable for optical clocks such as Ba^+, Sr^+, Ca^+, Hg^+, Yb^+, In^+, Tl^+, Ga^+, Al^+, and Lu^+, great progress have been made with $^{199}\mathrm{Hg}^+$ [13], $^{88}\mathrm{Sr}^+$ [14,15], $^{171}\mathrm{Yb}^+$ [16,17], $^{27}\mathrm{Al}^+$ [18], and$^{40}\mathrm{Ca}^+$ [19–22]. Table 10.1 shows the atomic parameters of optical atomic clocks based on $^{199}\mathrm{Hg}^+$,$^{88}\mathrm{Sr}^+$,$^{40}\mathrm{Ca}^+$,$^{138}\mathrm{Ba}^+$,$^{171}\mathrm{Yb}^+$,$^{115}\mathrm{In}^+$, and$^{27}\mathrm{Al}^+$. In recent years, the $^{88}\mathrm{Sr}^+$ optical clock developed by National Research Council of Canada has achieved a systematic uncertainty of 1.2×10^{-17} [15]. The $^{40}\mathrm{Ca}^+$ optical clock developed by Innovation Academy for Precision Measurement Science and Technology, Chinese Academy of Sciences, has achieved a systematic uncertainty of 1.3×10^{-17} [22]. The $^{199}\mathrm{Hg}^+$ optical clock developed by National Institute of standard and Technology (NIST) of USA has achieved a systematic uncertainty of 1.9×10^{-17} [13]. The $^{171}\mathrm{Yb}^+$ optical clock developed by Physikalisch-Technische Bundesanstalt (PTB) of Germany has achieved a systematic uncertainty of 3.0×10^{-18} [16]. The systematic uncertainty of the $^{27}\mathrm{Al}^+$ optical clock from NIST has reached 9.4×10^{-19} [18].

With the help of optical clocks, scientists can perform measurements at an unprecedented precision. By measuring the frequency ratio between two different optical clocks, one based on $^{199}\mathrm{Hg}^+$, the other based on $^{27}\mathrm{Al}^+$, an upper limit of the time-variance of the basic physical constant α was measured by NIST [13]. Recently, NPL and PTB used two clock transitions of $^{171}\mathrm{Yb}^+$ to make this measurement more precise [17]. NIST has also tested the theory of relativity via frequency comparisons between two optical clocks based on $^{27}\mathrm{Al}^+$ [23].

Here in Section 10.3.1, we will describe the basic concept of the ion-trapping technique. In Section 10.3.2, we will describe optical clocks referenced to the quadrupole transitions of ions, eg. $^{199}\mathrm{Hg}^+$, $^{88}\mathrm{Sr}^+$, $^{40}\mathrm{Ca}^+$, $^{138}\mathrm{Ba}^+$, and $^{171}\mathrm{Yb}^+$. In Section 10.3.3, we will describe optical clocks referenced to transitions other than the quadrupole transitions, eg., $^{171}\mathrm{Yb}^+$ (octupole transition) and $^{27}\mathrm{Al}^+$. In Section 10.3.4, the systematic frequency shifts and uncertainties of optical clocks based on single trapped ions will be given.

10.3.1 Ion Trapping

Ion traps adopt specific configuration of static electric and magnetic field or RF (radio frequency) field to confine ions in ultra-high vacuum environment, where the atomic system can be studied in an ideal isolated environment. Ions can stay there for a relatively long time, which is an ideal method to study atomic properties. Due to the invention of RF traps and the outstanding achievements with RF traps, Paul and Dehmelt won the Nobel Prize for Physics in 1989. Due to the achievement in terms of quantum manipulation and precision measurement with RF traps, Wineland won the Nobel Prize for Physics in 2012.

There are many different kinds of ion traps, such as Paul trap, Penning trap, and other special ion traps. The basic principles for these traps are similar. They are all based on specific configurations with DC electric and magnetic field or AC electric field. In the case of single ion trapping, scientists mainly adopt Paul trap [24] and Penning trap [25].

Since it requires weaker magnetic field during the detection of clock transitions of single ions, here we describe the principle of Paul trap in detail. Figure 10.2 shows a standard Paul

TABLE 10.1

Atomic Parameters of Optical Atomic Clocks Based on $^{199}\mathrm{Hg}^+$, $^{88}\mathrm{Sr}^+$, $^{40}\mathrm{Ca}^+$, $^{138}\mathrm{Ba}^+$, $^{171}\mathrm{Yb}^+$, $^{115}\mathrm{In}^+$, and $^{27}\mathrm{Al}^+$

Ion	Nuclear Spin (I)	Cooling Transition, Wavelength	Clock Transition, Wavelength	Clock Transition Natural Linewidth ()
$^{199}\mathrm{Hg}^+$	1/2	$^2S_{1/2} \leftrightarrow {}^2P_{1/2}$, 194 nm	$^2S_{1/2}\ (F = 0) \leftrightarrow {}^2D_{5/2}\ (F = 2)$, 282 nm	2 Hz
$^{88}\mathrm{Sr}^+$	0	$^2S_{1/2} \leftrightarrow {}^2P_{1/2}$, 422 nm	$^2S_{1/2} \leftrightarrow {}^2D_{5/2}$, 674 nm	0.4 Hz
$^{40}\mathrm{Ca}^+$	0	$^2S_{1/2} \leftrightarrow {}^2P_{1/2}$, 397 nm	$^2S_{1/2} \leftrightarrow {}^2D_{5/2}$, 729 nm	0.14 Hz
$^{138}\mathrm{Ba}^+$	0	$^2S_{1/2} \leftrightarrow {}^2P_{1/2}$, 493 nm	$^2S_{1/2} \leftrightarrow {}^2D_{3/2}$, 2051 nm	0.002 Hz
$^{171}\mathrm{Yb}^+$	1/2	$^2S_{1/2} \leftrightarrow {}^2P_{1/2}$, 370 nm	$^2S_{1/2}\ (F = 0) \leftrightarrow {}^2D_{3/2}\ (F = 2)$, 436 nm	22 Hz
			$^2S_{1/2}\ (F = 0) \leftrightarrow {}^2F_{7/2}(F = 3)$ 467 nm	10^{-9} Hz
$^{115}\mathrm{In}^+$	9/2	$^1S_0 \leftrightarrow {}^1P_1$, 159 nm	$^1S_0\ (F = 9/2) \leftrightarrow {}^3P_0\ (F = 9/2)$, 236.5 nm	0.8 Hz
$^{27}\mathrm{Al}^+$	5/2	$^1S_0 \leftrightarrow {}^1P_1$, 167 nm	$^1S_0\ (F = 5/2) \leftrightarrow {}^3P_0\ (F = 5/2)$, 267 nm	0.008 Hz

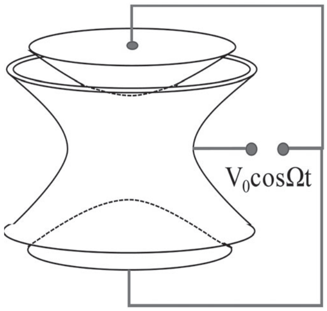


FIGURE 10.2 Configuration of a standard Paul trap.

trap, which consists of a ring electrode with rotational hyperbolic surface and two endcap electrodes with toroidal surface. The electric potential in the trap is generated by an alternating voltage as $U_0 - V_0\cos(\Omega t)$ between the ring electrode and two endcap electrodes, where U_0 is the DC voltage, V_0 is the RF voltage, and Ω is the frequency of the RF field.

Ions can only stay in the trap within a certain region, which is called stability region. The stability region depends on the mass of the ion, the dimension of the trap, and the frequency and amplitude of the RF field. Ion motion in the trap is a kind of harmonic oscillation driven by the RF trapping force. There are two kinds of motional frequencies, the secular motion at radial (axial) frequencies of $\omega_r\left(\omega_z\right)$ and micromotion with higher frequencies of Ω and smaller amplitudes compared to that of the secular motion.

One of the characteristics of the Paul trap is the trapping potential produced by the RF field which is associated with time, and somehow it can be described by the pseudo-potential model under certain conditions. Since there is no influence of strong magnetic field that could cause Zeeman shift, it has widespread applications in optical clocks.

10.3.2 Ion Optical Clocks Referenced to Quadrupole Transitions

Ions that have quadrupole transitions include ^{199}Hg$^+$, ^{171}Yb$^+$, ^{88}Sr$^+$, ^{40}Ca$^+$, ^{138}Ba$^+$, etc. These candidates have only one valence electron, which are similar to hydrogen, and their energy level structures are relatively simple and similar, as shown in Figure 10.3. The $^2S_{1/2} \leftrightarrow {}^2P_{1/2}$ electric dipole transition is suitable for laser cooling due to short lifetime of the $^2P_{1/2}$ state. The $^2D_{5/2}$ state is a metastable state with long lifetimes. Since the natural linewidth of the transition is inversely proportional to the lifetime of the upper level, the long lifetime (narrow natural linewidth) for the $^2S_{1/2} \leftrightarrow {}^2D_{5/2}$ transition makes it an ideal optical frequency reference. In the process of clock transition detection, in order to decrease the duty cycle, for ^{88}Sr$^+$ and ^{40}Ca$^+$, the time of the ion staying in the $^2D_{5/2}$ state can be reduced by introducing a quenching laser to pump the ion from the $^2D_{5/2}$ state to the $^2P_{3/2}$ state, and then the laser cooling is performed. For ^{199}Hg$^+$, the $^2S_{1/2}$ $(F = 1) \leftrightarrow {}^2P_{1/2}$ $(F = 0)$ transition is used for Doppler cooling and state preparation, and the $^2S_{1/2}$ $(F = 0) \leftrightarrow {}^2P_{1/2}$ $(F = 1)$ transition is used for state detection. The $^2S_{1/2}$ $(F = 0) \leftrightarrow {}^2D_{5/2}$ $(F = 2)$ transition provides the clock reference.

In a certain static magnetic field, the frequency of the clock transition can be shifted due to the first-order Zeeman effect. In addition, the electric field gradient can cause the quadrupole shift. Under normal circumstances, there are two main factors that cause frequency shift of the $^2S_{1/2} \leftrightarrow {}^2D_{5/2}$ transition. The $^2S_{1/2} \leftrightarrow {}^2D_{5/2}$ transition of ^{40}Ca$^+$ and ^{88}Sr$^+$ splits into 10 spectral lines in a static magnetic field, as shown in Figure 10.4. Six spectral lines shown in grey with numbers 1-6 in the figure can be chosen to probe and then average, by which the previously mentioned Zeeman shift and electric quadrupole shift can be effectively eliminated.

Once a single ion is trapped and laser cooled, the electric field and the magnetic field must be well controlled before probing the clock transition. The background magnetic field should also be well controlled. The magnetic field will introduce the Zeeman shift of the clock transition. The variation of the magnetic field broadens the observed spectrum and increases the systematic uncertainty. Experimentally, generally speaking, three pairs of coils were used to establish a controllable magnetic field in three dimensions. Moreover, it is necessary to minimize the micromotion in order to achieve a longer lifetime and a smaller frequency shift (the second-order Doppler shift and the Stark shift). Micromotion can be minimized by controlling the voltage of the ion trap by observing

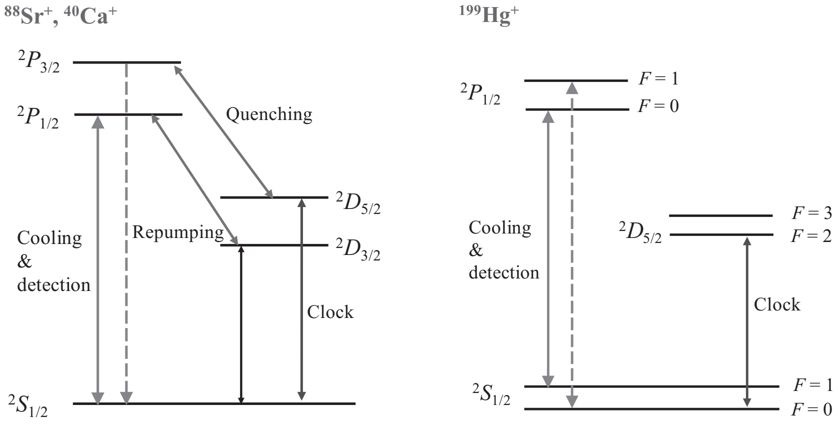


FIGURE 10.3 The lowest energy levels of ^{40}Ca$^+$, ^{88}Sr$^+$, and ^{199}Hg$^+$.

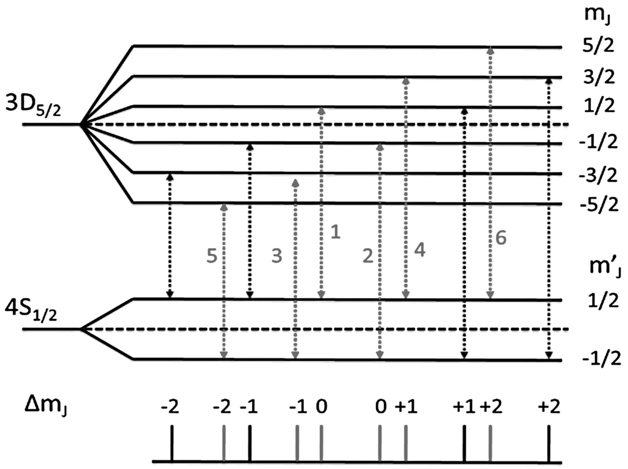


FIGURE 10.4 The sub-levels of the $^2S_{1/2} \leftrightarrow {}^2D_{5/2}$ transition for ^{40}Ca$^+$ and ^{88}Sr$^+$.

the ion position on a CCD camera and by monitoring the RF-photon correlation signal or the micromotion sidebands.

10.3.3 Ion Optical Clocks Referenced to the Transitions Other Than the Quadrupole Transitions

There are also ion candidates that take the transitions other than the quadrupole transitions as a frequency reference, such as ^{171}Yb$^+$ (octupole) [16], ^{115}In$^+$, and ^{27}Al$^+$ [18]. Ion clocks referenced to these ion candidates generally have more complex laser systems. The clock transitions of these ions are chosen for ultra-narrow natural transition linewidths or insensitivity to the blackbody radiation (BBR) shift.

As shown in Figure 10.5, ^{171}Yb$^+$ has two clock transitions: the 411 nm quadrupole and 467 nm octupole clock transitions. The two clock transitions of a single trapped ion can be used to test the time-variance of the fine structure constant α, taking advantage of a maximized reduction of systematic uncertainty [17,26]. The $^2S_{1/2}$ ($F = 0$) $\leftrightarrow$ $^2F_{7/2}$ ($F = 3$) electric octupole transition at 467 nm is highly forbidden, high-Q absorption brought about by the existence of a very-long-lived low lying metastable $^2F_{7/2}$ level, with a measured lifetime of ~ 3,700 days, corresponding to a natural spectral linewidth in the nHz range. The Q value of the octupole clock transition $^2S_{1/2}$ ($F = 0$) $\leftrightarrow$ $^2F_{7/2}$ ($F = 3$) is as large as 10^{24}, which is the largest among all known optical clocks. Although the natural linewidth of the ^{171}Yb$^+$ octupole clock transition is at the nHz level, the laser linewidth is usually at the 100 mHz level. Thus the probe laser intensity has to be increased, making the light shift very large. PTB has developed a method called hyper-Ramsey interrogation to reduce the light shift [27].

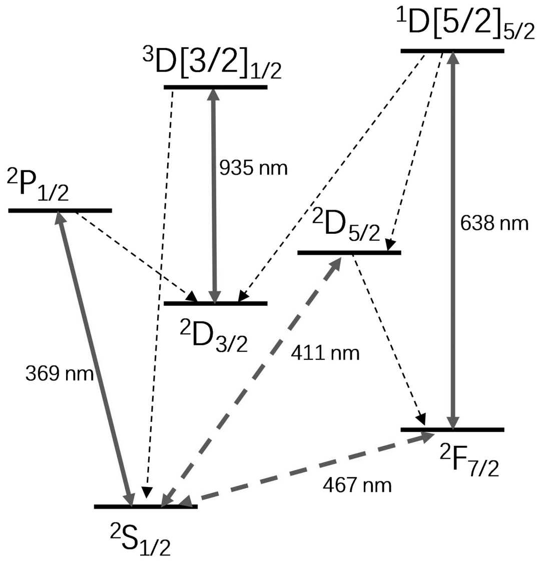


FIGURE 10.5 Partial term scheme for ^{171}Yb$^+$ showing 411 nm quadrupole and 467 nm octupole clock transitions.

The energy levels for ^{115}In$^+$ and ^{27}Al$^+$ are similar. As shown in Figure 10.6, the $^1S_0 \leftrightarrow {}^3P_0$ transition (clock transition) has no electric quadrupole shift, but there is still linear magnetic field dependence. The sensitivity of the clock transition of ^{115}In$^+$ and ^{27}Al$^+$ to blackbody Stark shift is the lowest known value among the ions and atoms studied. However, studying on ^{115}In$^+$ or ^{27}Al$^+$ is difficult due to the difficulty in building deep UV lasers. The cooling wavelength (the $^1S_0 \leftrightarrow {}^1P_1$ transition) for ^{27}Al$^+$ is located at 167 nm, which is very difficult to produce technically. Wineland's group at NIST has dealt with this problem in an interesting and novel way by making use of the techniques originally developed in their quantum information processing [28]. Two different ions are trapped simultaneously in the same trap, namely, a Be$^+$ ion (logic ion) and a ^{27}Al$^+$ ion (spectroscopy ion). The logic ion is laser-cooled using accessible cooling radiation at 313 nm, with the spectroscopy ion being sympathetically cooled by virtue of the combined motional mode of the ^{9}Be$^+$–^{27}Al$^+$ combination. The clock transition of ^{27}Al$^+$ at 267 nm is then probed. However, an additional difficulty is that there is no ^{27}Al$^+$ fluorescence available, which is used to observe clock transition quantum jump. This is also solved by using quantum information techniques to map the clock transition information of the ^{27}Al$^+$ ion back to the logic ion via motional states, adopting a blue-sideband (BSB) π pulse on the spectroscopy ion followed by a red-sideband (RSB) π pulse on the logic ion. In this way, the clock pulse information can be mapped back and read out in

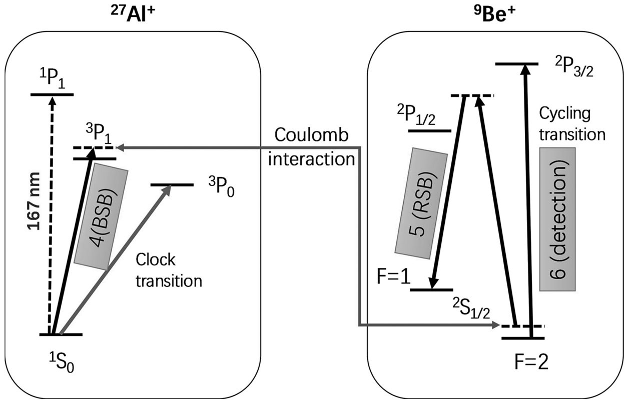


FIGURE 10.6 Transfer of the ^{27}Al$^+$ clock state to ^{9}Be$^+$ for detection.

TABLE 10.2

The Systematic Frequency Shifts and Their Uncertainties (u_B) of $^{199}Hg^+$, $^{88}Sr^+$, $^{40}Ca^+$, $^{171}Yb^+$, and $^{27}Al^+$ Optical Clocks

	$^{199}Hg^+$ [13]		**$^{88}Sr^+$ [15,29]**		**$^{40}Ca^+$ [30]**		**$^{171}Yb^+$ [16]**		**$^{27}Al^+$ [18]**	
Effect	**Shift**	**u_B**	**Shift**	**u_B**	**Shift**	**u_B**	**Shift**	**u_B**	**Shift**	**u_B**
BBR field (temperature)	0	0	557.6	11.6	842	12	−70.5	1.8	−3.05	0.42
BBR coefficient $\Delta\alpha_0$ including dynamic correction			/	0.8	/	<0.1				
Micromotion	−4	4	0	1	0	<0.1	−4.9	2.2	−4.58	0.59
Thermal motion	−3	3	−2.9	1.3	−1.5	0.2			−1.73	0.29
Quadrupole	0	10	0	≤1	0	<0.1	0	0.3	/	/
AC Stark	/	/	−4.5	2	0.1	0.1	0	1.1	0	0.2
Zeeman	−1130	11	<0.1	0.002	0	0.3	−40.4	0.6	−924.18	0.37
Servo	/	/	/	/	0	<0.1	0	0.5	/	/
Other	0	10	/	/	/	/	0	0.5	−0.06	0.35
Total	**−1137**	**19**	**550**	**12**	**828**	**13**	**−115.8**	**3.2**	−933.60	**0.94**

Shifts and uncertainties given are in fractional frequency units ($\Delta\nu/\nu$) in 10^{-18}.

~1 ms, causing little disadvantage in the timing budget for the full sequence.

The state detection ($^1S_0$ or $^3P_0$) is performed by repeating the following sequence:

1. Optical pumping π pulse ($^1S_0$, $F = 5/2$, $m_F = 3/2$) ↔ ($^3P_1$, $F = 7/2$, $m_F = 5/2$) (4 μs)
2. Sympathetic Doppler cooling (600 μs)
3. Ground-state cooling of the axial modes of the two-ion crystal (1 ms) $\bar{n} < 0.05$
4. BSB $^{27}Al^+$ π pulse (30 μs)
5. RSB $^9Be^+$ π pulse (7 μs)
6. Detection via the $^9Be^+$ cycling transition (200 μs)

10.3.4 Evaluation of Systematic Shifts

There is a variety of potential sources of systematic shifts associated with the clock transition in ion optical clocks. Generally speaking, systematic effects include the Doppler and Stark shifts caused by both the micromotion and the secular motion, the BBR Stark shift, the Zeeman shift, AC Stark shift (light shift), quadrupole shift, etc. For $^{88}Sr^+$,$^{199}Hg^+$, and $^{40}Ca^+$, the linear Zeeman shift, the quadruple shift, and the tensor part of the Stark shift can be cancelled out by averaging the frequencies of six spectral lines. For $^{88}Sr^+$ and ^{40}Ca, at a certain ion-trapping RF frequency, the so-called magic-trapping frequency, the second-order Doppler shift and the scalar part of the Stark shift caused by both the micromotion and the secular motion would also be cancelled out [29,30]. With the help of high-quality magnetic shielding, the clock can run under weak and stable magnetic field, making the second-order Zeeman shift below 1×10^{-18}. All the lasers except the clock laser are blocked by mechanical shutters during spectroscopy, largely eliminating the light shift. Theoretically, the frequency shift due to the clock laser is very small when using long interrogation time. Experimentally, this shift can be precisely measured. For $^{88}Sr^+$ and $^{40}Ca^+$ clocks, the total uncertainty is limited by the BBR shift, which highly depends on the measurement precision of environmental temperature. For $^{171}Yb^+$ and $^{27}Al^+$ clocks, the total uncertainty is limited by the motion effects. For the Hg$^+$ clock, the total uncertainty is limited by the instability of magnetic field. Table 10.2 summarizes the systematic frequency shifts evaluation for $^{199}Hg^+$, $^{88}Sr^+$, $^{40}Ca^+$, $^{171}Yb^+$, and $^{27}Al^+$ optical clocks.

10.4 Optical Clocks Based on Neutral Atoms

Compared to single trapped ion optical atomic clocks, optical atomic clocks based on neutral atoms have the advantage of large atomic number, which gives enhanced clock signals and thus lower quantum-projection-noise-limited frequency instabilities. With the advance of improved ultra-stable lasers used as local oscillators, optical atomic clocks based on neutral atoms make breakthrough in frequency stability and accuracy.

The choice of atomic species for optical clocks is based on two criterions. (i) It has a transition with a large quality factor ($Q = \nu/\Delta\nu$, where ν and $\Delta\nu$ are the frequency and linewidth of the transition, respectively) and its frequency ν is insensitive to external fields. (ii) The atoms can be easily laser-cooled and confined to remove the first-order Doppler effect in order to obtain a spectrum with a linewidth approaching to the natural linewidth of the transition.

For alkaline earth atoms with two valence electrons, the electronic structure is arranged symmetrically into singlet states ($S = 1$) and triplet states ($S = 3$), as shown in Figure 10.7. Similar to the alkaline earth metals, the lanthanide ytterbium and the transition metals zinc, cadmium, and mercury, and the synthetically prepared actinide nobelium have two valence electrons. Strong dipole-allowed transitions exist when satisfying $\Delta S = 0$, e.g. $^1S \leftrightarrow {}^1P$,$^3P \leftrightarrow {}^3S$, and $^3P \leftrightarrow {}^3D$. These transitions are well suited for laser-cooling, trapping, and repumping. By contrast, transitions between singlet and triplet states are generally forbidden due to $\Delta S \neq 0$. However, the $^1S_0 \leftrightarrow {}^3P_1$ transition and the $^1S_0 \leftrightarrow {}^3P_0$ transition of fermionic isotopes (odd-mass-number isotopes) are weakly allowed as a result of $L{\cdot}S$ interaction or hyperfine interaction among the excited states. The $^3P_0$ state has a very weak dipole coupling to

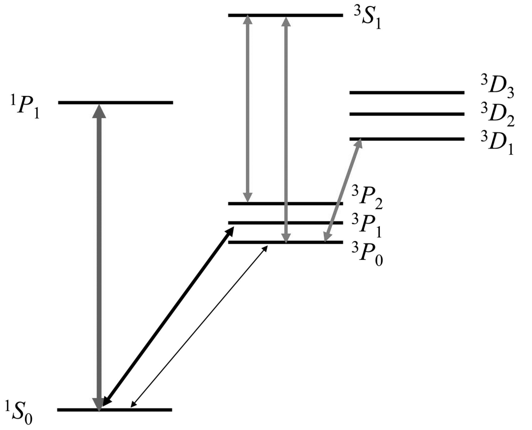


FIGURE 10.7 Energy level diagram related to laser cooling, spectroscopy, and repumping.

$^1S_0$, and the natural linewidth $\Delta\nu_n$ of the $^1S_0 \leftrightarrow {}^3P_0$ transition is usually below 1 Hz, which satisfies the first criterion as an optical clock. Among atoms with two valence electrons, Sr, Yb, and Hg are popularly exploited. We will introduce optical atomic clocks based on Sr, Yb, and Hg in the following text.

For Sr, the popular isotope used in optical atomic clocks is ^{87}Sr, which has an abundance of 7% and nuclear spin of $I = 9/2$. By contrast, ^{171}Yb and ^{199}Hg have the smallest nuclear spin of $I = 1/2$, which simplifies the m_F states of $^1S_0$ or $^3P_0$ into $m_F = \pm 1/2$ two states. Some parameters of Sr, Yb, and Hg are listed in Table 10.3.

10.4.1 Preparation of Neutral Atoms for Interrogation

At room temperature, Sr and Yb are solid while Hg is liquid. In order to have big enough atomic beam flux, Sr and Yb are heated in an effusive oven at nearly 600°C and 450°C, respectively. Therefore, atoms of Sr or Yb fly out with a most probable velocity more than 300 m s^{-1}. Those atoms are slowed by repeatedly absorbing red-detuned photons propagating from the opposite direction in order to be efficiently captured in a magneto-optical trap (MOT). Strong transitions such as $^1S_0 \leftrightarrow {}^1P_1$ are usually chosen to cool and trap the atoms. The $^1S_0 \leftrightarrow {}^1P_1$ transition typically has a natural linewidth $\Delta\nu_n$ more than 10 MHz, allowing fast photon scattering for efficiently cooling. However, the coldest temperature achieved with those transitions is limited to nearly 1 mK according to the following equation (Doppler cooling limit):

$$T_D = \frac{h \times \Delta\nu_n}{2k_B} \tag{10.2}$$

where h and k_B are Planck constant and Boltzmann's constant, respectively. According to this equation, the achieved temperature using atomic transitions with relatively narrower linewidths can be lower, which is the motivation of the second-stage laser cooling. The transitions for the second-stage cooling of Sr and Yb are usually based on the $^1S_0 \leftrightarrow {}^3P_1$ transition, which has Doppler-limited cooling temperatures down to ~1 μK.

In contrast, Hg is liquid at room temperature. Atoms of Hg need to be cooled down to −40°C [31]. Since the most probable velocity of Hg at −40°C is less than 100 m s^{-1}, Hg atoms are directly captured in an MOT. Since the wavelength of the $^1S_0 \leftrightarrow {}^1P_1$ transition of Hg is below 200 nm and the laser source at this wavelength is unavailable, Hg atoms are directly cooled and trapped with the $^1S_0 \leftrightarrow {}^3P_1$ transition.

Even the above atoms are cooled down to ~1 μK, the Doppler-broadened spectroscopy has a linewidth of tens of kHz, much larger than the natural linewidth of the clock transitions below 1 Hz. Instead of merely reducing the atomic temperature below 10^{-16} K for a Doppler-broadened spectral linewidth at the Hz level, it is more efficient to reduce the first-order Doppler effect by tightly confining the atoms in optical lattice sites to the Lamb-Dicke regime, where the atoms are

TABLE 10.3

Atomic Parameters of Optical Atomic Clocks Based on Sr, Yb, and Hg

	^{87}Sr	**^{171}Yb**	**^{199}Hg**
Nuclear spin (I)	9/2	1/2	1/2
Natural abundance	7.02%	14.3%	16.9%
Magnetic moment	−1.0924	+0.4919	+0.5059
Strong cooling	$^1S_0 \leftrightarrow {}^1P_1$		
λ, $\Delta\nu_n$, T_D	460.9 nm, 32 MHz, 770 μK	398.9 nm, 29 MHz, 700 μK	184.9 nm, 120 MHz (unused)
Narrow cooling	$^1S_0 \leftrightarrow {}^3P_1$		
λ, $\Delta\nu_n$, T_D	689.3 nm, 7.5 kHz, 0.2 μK	555.8 nm, 180 kHz, 4 μK	253.7 nm, 1.3 MHz, 30 μK
Clock transition	$^1S_0 \leftrightarrow {}^3P_0$		
λ, $\Delta\nu_n$	698.4 nm, 1 mHz	578.4 nm, 10 mHz	265.5 nm, 100 mHz
Repump I	$^3P_2 \leftrightarrow {}^3S_1$		
λ, $\Delta\nu_n$	707.2 nm, 7.3 MHz	770.2 nm, 5.9 MHz	546.1 nm, 7.8 MHz
	$^3P_0 \leftrightarrow {}^3S_1$		
	679.3 nm, 1.4 MHz	649.1 nm, 1.5 MHz	404.6 nm,3.3 MHz
Repump II	$^3P_0 \leftrightarrow {}^3D_1$		
λ, $\Delta\nu_n$	2603 nm, 56 kHz	1388 nm, 420 kHz	296.8 nm, 7.3 MHz
Magic wavelength	813.4 nm	759.4 nm	362.57 nm

confined in a region much smaller than the wavelength of the probe light [32]. By tightly confining the atoms, the photon-recoil effect is also reduced. Usually the intensity of the lattice light is on the order of 10^3 W cm^{-2} to make the optical trap depth high enough in order to load more atoms from the MOT. The lattice light with high intensity also shifts the energy levels and deteriorates the accuracy of optical atomic clocks. By setting the wavelength of the lattice light to the "magic wavelength", as shown in Table 10.1, the shifts of the $^1S_0$ and $^3P_0$ states can be the same, resulting in an unshifted clock transition [33].

10.4.2 Spectroscopy and Laser Locking

Due to nuclear spin, the energy levels of $^1S_0$ and $^3P_0$ have a magnetic state degeneracy of $2I + 1$. In small or negligible background magnetic fields, these nuclear spin states are degenerate, and atoms are distributed among these nuclear spin states. Since each magnetic state has different Zeeman and Stark shift, this mixed atomic population is not ideal for high-accuracy spectroscopy. Thus, the atoms are usually prepared in a certain spin state by purposely applying a magnetic field to split the nuclear spin state degeneracy and optical pumping on the $^1S_0 \leftrightarrow {}^3P_1$ transition to transfer the atoms to a single ground state.

During the spectroscopy of the $^1S_0 \leftrightarrow {}^3P_0$ clock transition, a small magnetic field is applied in order to separate the m_F states. The polarization of the probe light and the lattice light is aligned with the magnetic field. Usually a π pulse resonant with the clock transition is applied for a time during T_{int}, yielding a Fourier-limited line shape of $0.89/T_{int}$. Instead of using a fluorescence signal from the excited state of an absorption signal, an electron-shelving technique is usually employed to efficiently observe the weak transition [34]. As a result of partial excitation of the atomic sample by the probe light, atoms not shelved into the excited state can be measured by observing fluorescence from a pulse of light resonant on the $^1S_0 \leftrightarrow {}^1P_1$ transition (detection light). This scattered fluorescence is then collected on a photomultiplier tube. In order to reduce the fluctuation of atomic number and background noise, and to obtain a better signal-to-noise ratio (SNR) of the clock transition, the fraction of excited atoms is detected by applying a sequence of three pulses of detection light [35]. The first pulse measures and heats away all atoms remaining in the $^1S_0$ state after spectroscopy, N_g. The other two pulses measure the background signal N_b due to fluorescence from the thermal atomic beam and the atoms N_e that were excited to the $^3P_0$ state by pumping the atoms back to the ground state via either $^3P_0 \leftrightarrow {}^3S_1$ and $^3P_2 \leftrightarrow {}^3S_1$ or the $^3P_0 \leftrightarrow {}^3D_1$ transition. Then, the excitation rate is calculated as $R = (N_e - N_b)/(N_e + N_g - 2N_b)$.

To lock the probe laser to the atomic transition, usually the probe laser frequency is stabilized to the average of the two π transitions between the nuclear Zeeman sub-states [36], as shown in Figure 10.8. At first, the laser frequency is set to the low frequency half maximum of the first peak (point *a*), and spectroscopy is performed to detect the excitation rate R_a. In the second measurement cycle, the probe laser frequency is set to *b* in the figure, and the excitation rate R_b is detected. The excitation rates at *a* and *b* are computed to generate the first error signal $\Delta_1 = R_a - R_b$. When $R_a = R_b$, it indicates that the laser is centred at the first peak. Otherwise, the laser frequency needs to be adjusted by feeding the error signal to a servo to feed back to any laser frequency shifter such as AOMs. Similarly, in the third and fourth measurements, the excitation rates are detected at *c* and *d*, and the laser frequency is adjusted to be centred at the second peak. For every four measurements, the outputs of the two servos are averaged and fed to a third servo to adjust the frequency of AOMs.

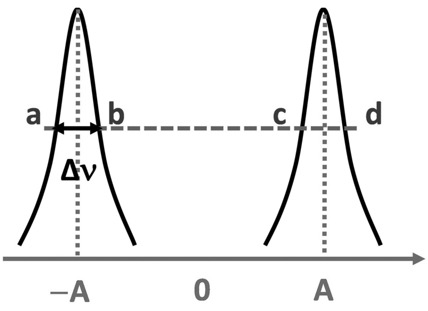


FIGURE 10.8 Schematic diagram for laser stabilization to the average of two π clock transitions under a bias magnetic field.

10.4.3 Systematic Evaluations

As an accurate frequency standard, the output frequency of optical clocks should be insensitive to external fields and other environmental parameters. For optical clocks based on neutral atoms, they have the advantage of higher SNR and thus better frequency stability. At the same time, their frequencies will be influenced by strong lattice light and atomic collision, compared to optical frequency standards based on a single ion.

The energy levels of atoms experience Stark shift in the presence of electric field E produced by the lattice light, $U = -E_i \alpha_{ij} E_j / 2$. The polarizability α can be written as the sum of scalar, vector, and tensor polarizabilities. When the wavelength of the lattice light is kept at the magic wavelength with an uncertainty less than 500 kHz, the scalar Stark shift can be controlled at the 10^{-18} level. The vector Stark shift arises from imperfect linear polarization of the one-dimensional lattice light. Since the vector Stark shift is proportional to m_F magnetic sub-states, it can be cancelled by averaging the opposite m_F magnetic sub-states, as shown in Figure 10.8. Unlike the vector Stark shift, the tensor Stark shift has $3\,m_F^2 - F\,(F + 1)$ dependence. Fortunately, the tensor Stark shift is zero for Yb and Hg since $F = 1/2$ and $m_F = \pm 1/2$, and it is small for Sr, which can be cancelled together with the scalar Stark shift by properly setting the lattice frequency and depth. Besides the above Stark shift due to electric-dipole ($E1$) couplings, magnetic dipole ($M1$), electric quadrupole ($E2$), and two-photon can also lead to small Stark shift. Experiments show that for Yb atoms trapped in lattice with a depth of $50E_r$, where E_r is the recoil energy, the systematic uncertainty due to the above multipole couplings is less than 3×10^{-19} [3].

In most optical atomic clocks based on neutral atoms, atoms are trapped in one-dimensional lattice. There are several atoms in one lattice site, which may cause collision and shift the energy levels. Since fermions are usually used, identical fermions cannot interact with each other through *s*-wave. However, if the atoms are not prepared in identical quantum states, the atoms are not distinguishable and will be interacting with each other via *s*-wave collision. By effectively optical pumping Yb atoms into one of the m_F magnetic sub-states with 99.5% efficiency, the *s*-wave collision shift can be largely suppressed to 3×10^{-19} [3]. Secondly, finite atomic temperatures prevent the *p*-wave collision from being completely suppressed. Such collision can be effectively reduced by reducing the atomic temperature below the *p*-wave barrier. By evaluating the density-dependent collision shift at a range of atomic numbers and trap depths, the collision shift can be evaluated [37,38].

Since the atoms are surrounded by matter with non-zero temperature, which radiates electromagnetic wave. BBR shifts the energy levels of the atoms via the Stark shift. Since the peak of the BBR spectrum is far red-detuned from clock transitions, the BBR shift can be modelled as a DC Stark shift with a small dynamic correction η accounting for non-static contribution to the shift [39]. The Stark shift of the clock transition due to BBR can be written as

$$\begin{aligned}\Delta\upsilon &= -\frac{1}{2}\left(\alpha_{\mathrm{e}}^{(0)} - \alpha_{\mathrm{g}}^{(0)}\right)\left\langle E^2\right\rangle_T\left[1+\eta(T)\right]\\ &= -\frac{1}{2}\alpha_{\mathrm{clock}}^{(0)}\left\langle E^2\right\rangle_T\left[1+\eta(T)\right] \qquad (10.3)\end{aligned}$$

where $\alpha_{\mathrm{g,\,e}}^{(0)}$ are the static dipole polarizabilities of ground and excited clock states, $\left\langle E^2\right\rangle_T \approx (8.319\ \mathrm{V/cm})^2(T/300\ \mathrm{K})^4$ is the BBR intensity, T is the temperature of the blackbody, and $\alpha_{\mathrm{clock}}^{(0)}$ is the static dipole polarizability of the clock transition. The static dipole polarizabilities $\alpha_{\mathrm{clock}}^{(0)}$ of Yb and Sr are measured, respectively, to be $3.62612 \times 10^{-2}\,\mathrm{Hz}\ (\mathrm{V\ cm^{-1}})^{-2}$ and $6.15558 \times 10^{-2}\,\mathrm{Hz}\ (\mathrm{V\ cm^{-1}})^{-2}$ with an uncertainty at the 20 ppm level [40,41]. The dynamic corrections are determined to be $\eta_{\mathrm{Yb}}(T) = 0.0173\ (T/300\ \mathrm{K})^2 + 0.0006(T/300\ \mathrm{K})^4$ and $\eta_{\mathrm{Sr}}(T) = 0.069\ (T/300\ \mathrm{K})^2 + O(T/300\ \mathrm{K})^4$ with uncertainties less than 3% [41,42]. Therefore, at room temperature $T = 300$ K, the BBR shift for Sr and Yb is 2.4 Hz and 1.3 Hz, respectively. When the temperature is measured with an uncertainty of 1 K at room temperature, the uncertainty arisen from the BBR shift is 30 mHz (7×10^{-17}) and 18 mHz (3.5×10^{-17}) for Sr and Yb, respectively. In order to reduce the BBR shift to 10^{-18} level, environmental temperature should be controlled and measured with an uncertainty within 0.1 K. While for Hg, its susceptibility to the BBR is an order of magnitude smaller than that of Sr and Yb. When environmental temperature is measured with an uncertainty of 3 K at room temperature, the uncertainty of the BBR shift is 34 mHz (3×10^{-17}) [43].

In a magnetic field B, the frequency of a clock transition (π-transitions) is shifted on the first order as $\Delta\nu_{\mathrm{B1}} = -m_{\mathrm{F}}\delta g\mu_{\mathrm{B}}B/h$, where δg is the difference in the g-factors of the excited and ground states and μ_{B} is Bohr magneton. For Sr and Yb, $\delta g\mu_{\mathrm{B}}/h$ is determined experimentally to be −1.084(4) Hz/μT and −1.91(7) Hz/μT, respectively [44,45]. Since two nuclear spin states are interrogated and an average between them is taken, the first-order Zeeman shift can be cancelled completely. However, the second-order Zeeman shift is common to both π-transitions, and thus it cannot be cancelled by averaging. The second-order Zeeman shift can be determined by $\Delta\nu_{\mathrm{B2}} = \beta B^2$, where β is -2.33×10^{-5} Hz/μT², -6.2×10^{-6} Hz/μT²and 1.6×10^{-6} Hz/μT² for Sr, Yb, and Hg, respectively [35,43,46]. The measured splitting of two π-transitions yields the magnetic field to determine the second-order Zeeman shift.

Other systematic uncertainties come from DC Stark shift due to stray DC electric field, AC Stark shift due to the probe light, tunnelling shift arisen from tunnelling of atoms between lattice sites, and Doppler shift due to the motion of lattice sites relative to the probe light. These shifts can be controlled to be less than 1×10^{-18}. Nowadays, the most accurate optical atomic clocks have achieved systematic frequency-shift uncertainties of $(1–2) \times 10^{-18}$ [2, 3].

10.5 Absolute Frequency Measurement and Optical Frequency Synthesis Using Femtosecond Combs

The frequency stabilities and the systematic frequency-shift uncertainties provided by optical atomic clocks have reached 10^{-18}. The unprecedented accuracy is fostering a revolution in science and technology: tests of fundamental physics, geodesy, searching for dark matter, the redefinition of the SI second, and applications based on the next generation of clocks including GNSS, GPS, radar, radio-astronomy, and communication networks [47]. However, the oscillation of optical waves, $>10^{14}$ oscillations in 1 second, is too fast to be counted for any electronic device. As optical frequency standards, their frequency values must be measured relative to the caesium primary frequency standard at about 9.2 GHz, as this defines the SI second. It is not an easy task to link the large frequency gap between the optical region and the microwave region. Early attempts such as frequency measurement chains based on a large number of intermediate oscillators with different frequencies that were compared and phase-locked by harmonic mixing in non-linear elements and beat frequency measurements were too complex and difficult to operate [48]. In 1999, a revolutionary technology using Kerr-lens mode-locked femtosecond Ti:sapphire lasers emerged [49]. The mode-locked femtosecond laser can emit coherent optical waves, whose frequencies are discrete and equally spaced with frequency intervals defined by the laser repetition rate f_{r}. In the frequency domain, the laser spectrum looks like a comb, named optical frequency comb. Later, the spectrum of the femtosecond laser was extended using photonic crystal fibre or microstructure fibre from a few tens of THz to an octave, e.g. 530–1060 nm [50]. It enables the detection and stabilization of the carrier-envelope offset frequency (f_0) of the comb [51].

In the frequency domain, as shown in Figure 10.9, the frequency of each comb tooth is

$$f_{\mathrm{n}} = nf_{\mathrm{r}} + f_0 \qquad (10.4)$$

where n is an integer. The carrier-envelope offset frequency f_0 is less than f_{r}. The repetition rate f_{r} is determined by the round trip time of the light inside the laser cavity as $f_{\mathrm{r}} = 1/\tau$.

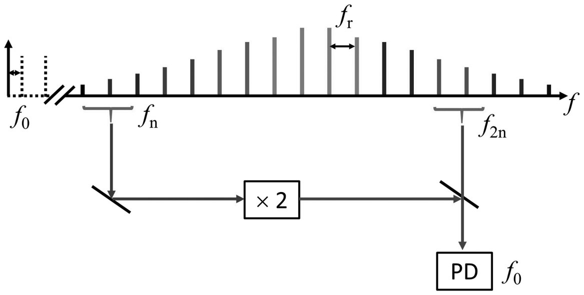


FIGURE 10.9 1*f*-2*f* self-referenced optical frequency comb.

Different types of combs have different repetition rates. For those based on mode-locked Ti:sapphire lasers, f_r can be several hundreds of MHz to several GHz. For fibre-based optical frequency combs, f_r is usually several tens of MHz to 500 MHz. For chip-scale micro-combs, f_r can be on the order of 10^{10} Hz to 10^{12} Hz. When the frequency of f_r is within the detection bandwidth of a photodetector (PD), it can be easily detected by detecting a portion of the laser light. Then, it can be stabilized by tuning the length of the laser cavity.

Measurement of the carrier-envelope offset frequency is usually done with *1f-2f* self-referencing technique [51]. As shown in Figure 10.9, light at the low-frequency region of the broadened spectrum, e.g. 1060 nm, is frequency-doubled in a non-linear crystal. The second-harmonic generation heterodyne beats against the high-frequency region of the comb at the same wavelength, e.g. 530 nm. The beating signal is detected on a PD, which is

$$2\times f_n - f_{2n} = 2\times\left(nf_r + f_0\right) - \left(2nf_r + f_0\right) = f_0 \qquad (10.5)$$

Then, the detected f_0 signal is stabilized by controlling the dispersion inside the laser cavity. The actuator can be prisms or pump power to tune the Kerr phase shift. The breakthrough which led to a practical frequency comb was the development of technology for stabilizing the carrier-envelope offset frequency.

10.5.1 Absolute Optical Frequency Measurement

If an unknown optical frequency, f_{cw}, needs to be measured relative to a microwave frequency standard, e.g. the Cs primary frequency standard, H maser, or GPS-disciplined Rb clock, the laser light heterodyne beats against the comb light on a PD to obtain the beating frequency f_b between f_{cw} and the nearby comb tooth f_n, as depicted in Figure 10.10. The frequency values of f_r, f_0, and f_b are synchronously measured on RF counters referenced to the same microwave frequency standard. Then, the value of f_{cw} can be derived from

$$f_{cw} = nf_r \pm f_0 \pm f_b \qquad (10.6)$$

If a high-precision wavemeter is available, the frequency uncertainty of f_{cw} can be coarsely determined with this wavemeter. Thus, the ambiguity of signs and the comb number n in equation (10.6) can be determined.

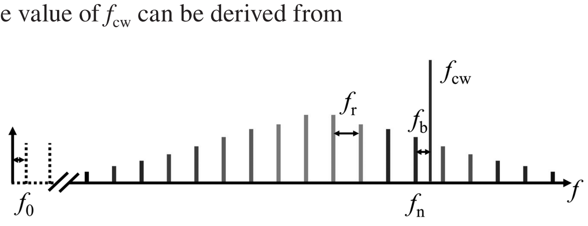


FIGURE 10.10 Absolute optical frequency measurement.

To reduce the comb frequency noise in measuring the value of f_{cw}, the optical frequency comb, especially f_0 and f_r, is stabilized to a microwave frequency standard using phase-locked loops. In each phase-locked loop, the signal of f_0 (f_r) compares against an RF signal f_{RF1} (f_{RF2}) synthesized from the microwave frequency standard to generate an error signal, which is then sent to a servo to stabilizing f_0 (f_r) to f_{RF1} (f_{RF2}). By stabilizing the optical frequency comb, the frequency instability of each comb tooth is roughly at the same performance of the microwave frequency standards, e.g. 10^{-11} at 1 s averaging time for Rb clock and 10^{-13} for H maser. By doing this, the uncertainty of the optical frequency measurement limited by the RF frequency standard may be achieved in a shorter time, compared to that using a free-running optical frequency comb.

10.5.2 Optical Frequency Ratio Measurement

Towards the redefinition of the SI second, the following procedures should be made according to Ref. [6]. "(i) At least three different optical clocks (either in different laboratories, or of different species) have demonstrated validated uncertainties of about two orders of magnitude better than the best Cs atomic clocks of time. (ii) At least three independent measurements of at least one optical clock from the first procedure have been compared in different institutes (with e.g. $\Delta\nu/\nu < 5 \times 10^{-18}$) either by transportable clocks, advanced links, or frequency ratio closure. (iii) Three independent measurements of the optical frequency standards of the first procedure with three independent Cs primary clocks have been performed, where the measurements are limited essentially by the uncertainty of these Cs fountain clocks (with e.g. $\Delta\nu/\nu < 3 \times 10^{-16}$). (iv) Optical clocks (secondary representations of the second) contribute regularly to TAI. (v) Optical frequency ratios between a few (at least five) other optical frequency standards have been performed; each ratio measured at least twice by independent laboratories and agreement was found to better than, e.g. $\Delta\nu/\nu < 5 \times 10^{-18}$". Therefore, the measurements of frequency ratios between optical atomic clocks are critical to the redefinition of the SI second.

To measure the frequency ratio between two cw lasers or optical atomic clocks, f_{cw1} and f_{cw2}, the beating signals, f_{b1} and f_{b2}, between the cw lasers and their nearby comb teeth must be detected, and their frequencies are measured on RF counters referenced to the same RF frequency standard. Similar to absolute optical frequency measurement, the frequency ratio between two lasers can be determined by

$$r = \frac{f_{cw1}}{f_{cw2}} = \frac{n_1 f_r \pm f_0 \pm f_{b1}}{n_2 f_r \pm f_0 \pm f_{b2}} \qquad (10.7)$$

Since f_r, f_0, f_{b1}, and f_{b2} are all measured relative to the same stable RF reference frequency, the absolute frequency of the RF reference does not affect the ratio. Any error in the reference frequency causes all terms on the right-hand side of equation (10.7) to scale equally, thus leaving the ratio r unchanged. If the comb is frequency-stabilized to one of the cw lasers with frequency instability better than 1×10^{-15} at 1 s averaging time, the measurement noise will be decreased accordingly, which reduces the measurement time and measurement uncertainty as well. Since various counters and synthesizers involved in the measurement track the RF reference to some extent, the measurement uncertainty using this method is limited to 10^{-19} [13].

One of the methods to reduce the measurement noise caused by the optical frequency comb is called transfer oscillator scheme [52]. In the transfer oscillator scheme, shown in Figure 10.11, the frequency noise of the comb appearing in f_r and f_0 is subtracted from the beating signals of f_{b1} and f_{b2} by using RF mixers and direct digital synthesizers (DDSs). The divisors of the DDSs satisfy $K_1/N_1=K_2/N_2$, where N_1 and N_2 are integers related to comb tooth number. Afterwards, a comb-noise-free signal is obtained as

$$\Delta = \frac{f_{\mathrm{cw1}}}{K_1} - \frac{f_{\mathrm{cw2}}}{K_2} \tag{10.8}$$

Therefore, the frequency ratio between the cw lasers is

$$r = \frac{f_{\mathrm{cw1}}}{f_{\mathrm{cw2}}} = \frac{K_1\left(\Delta + \dfrac{f_{\mathrm{cw2}}}{K_2}\right)}{f_{\mathrm{cw2}}} = \frac{K_1}{K_2} + K_1\frac{\Delta}{f_{\mathrm{cw2}}} \tag{10.9}$$

Since $\dfrac{\Delta}{f_{cw2}}$ is usually below 10^{-7}, Δ and f_{cw2} need to be measured to fractional uncertainties of 10^{-11} in order to measure the optical frequency ratio with an uncertainty below 10^{-18}. In conjunction with synchronous counting to reduce the noise from the common RF time-base, a fractional uncertainty of optical frequency ratio measurement has been demonstrated below 3×10^{-21} over a data collection period of 10^5 s [10].

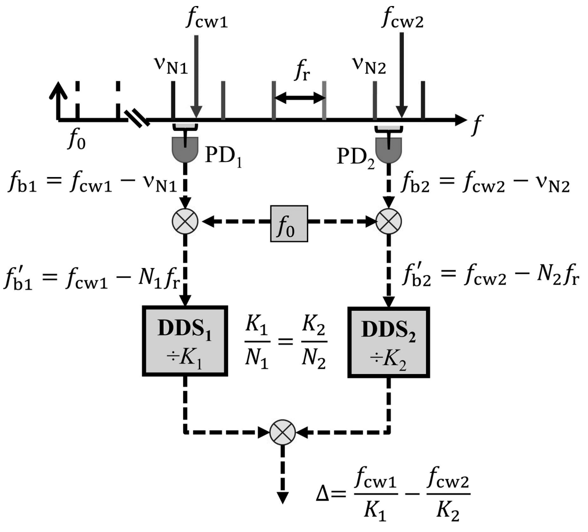

f_{cw1}
f_{cw2}
ν_{N1}
f_{r}
ν_{N2}
f
f_0
PD$_1$
PD$_2$
$f_{\mathrm{b1}} = f_{\mathrm{cw1}} - \nu_{\mathrm{N1}}$
$f_{\mathrm{b2}} = f_{\mathrm{cw2}} - \nu_{\mathrm{N2}}$
f_0
$f'_{\mathrm{b1}} = f_{\mathrm{cw1}} - N_1 f_{\mathrm{r}}$
$f'_{\mathrm{b2}} = f_{\mathrm{cw2}} - N_2 f_{\mathrm{r}}$
DDS$_1$ $\div K_1$
$\frac{K_1}{N_1} = \frac{K_2}{N_2}$
DDS$_2$ $\div K_2$
$\Delta = \frac{f_{\mathrm{cw1}}}{K_1} - \frac{f_{\mathrm{cw2}}}{K_2}$

FIGURE 10.11 The schematic diagram of the transfer oscillator scheme for removing the frequency noise induced by the optical frequency comb.

10.5.3 Optical Frequency Synthesis

Optical atomic clocks provide unprecedented frequency stability and uncertainty. For practical applications of optical atomic clocks, their frequencies need to be synthesized to any desired optical frequency or even to the microwave region.

To transfer the coherence from an optical atomic clock to a desired optical frequency, usually an optical frequency comb is firstly frequency-stabilized to the optical atomic clock. The carrier-envelope offset frequency of the comb and the beating signal between the optical atomic clock and the nearby comb tooth are phase-locked to the same RF frequency standard by feeding back to the dispersion and cavity length of the comb laser. Thereby, the comb inherits the frequency stability and accuracy from the optical atomic clock. Then, a laser at the desired frequency is phase-locked to the frequency-stabilized comb by phase-locking its beating signal against its nearby comb tooth to the RF frequency standard. Since the frequencies of RF signals are 10^7 smaller than the optical frequency, the frequency instabilities of RF signals can be 10^{-11} when synthesizing the signal of optical atomic clocks with frequency instabilities of 10^{-18}.

Alternatively, an optical frequency comb is used as a transfer oscillator to transfer the coherence from one laser to another. This scheme is frequently used when some optical frequency combs cannot faithfully transfer the coherence of the optical atomic clock to every comb tooth due to limited servo bandwidth. In this case, the comb is coarsely frequency-stabilized to keep all the signals within the bandwidth of RF band-pass filters, which are usually used for obtaining signals with better SNRs. The error signal for locking the laser light at the desired frequency to the optical atomic clock is generated as shown in Figure 10.11. To realize precise frequency tuning of the synthesized laser light, a tunable RF signal can be mixed with f'_{b2} in Figure 10.11.

In order to remove the frequency noise from RF signals, which is irrelevant to the optical atomic clock, a self-referenced RF signal might be better synthesized directly from the optical clock using the same optical frequency comb [9]. The frequency of the self-referenced RF signal is $f_{\mathrm{RF}} = kf_{\mathrm{c}}$, where f_{c} is the frequency of the optical atomic clock and k is a divisor. By doing this, the desired laser frequency f_{d} is only related to the optical atomic clock with a ratio of $r = f_{\mathrm{d}}/f_{\mathrm{c}}$. This is similar to RF frequency dividers but in the optical region, which is called optical frequency divider. Nowadays, optical frequency divider can convert optical frequencies with an uncertainty of 1.4×10^{-21} and a frequency instability better than 1×10^{-18} at 1 second, two orders of magnitude smaller than the most accurate optical atomic clocks and the most stable laser light.

10.6 Internationally Recommended Optical Frequency Standards and Optical Clocks

Recent development of optical atomic clocks has shown that the frequency uncertainties of optical atomic clocks have surpassed the primary frequency standard. One of these clocks could provide the basis for a future definition of the SI second.

TABLE 10.4

“Secondary Representations of the Second” of 2017

Frequency (Hz)	Fractional Uncertainty	Transition
6834 682 610.904 3126	6×10^{-16}	^{87}Rb ground state hfs
429 228 004 229 873.0	4×10^{-16}	^{87}Sr, $5s^2\ ^1S_0 \leftrightarrow 5s5p\ ^3P_0$
444 779 044 095 486.5	1.5×10^{-15}	^{88}Sr$^+$, $5s\ ^2S_{1/2} \leftrightarrow 4d\ ^2D_{5/2}$
518 295 836 590 863.6	5×10^{-16}	^{171}Yb, $6s^2\ ^1S_0 \leftrightarrow 6s6p\ ^3P_0$
642 121 496 772 645.0	6×10^{-16}	^{171}Yb$^+$, $^2S_{1/2} \leftrightarrow {}^2F_{7/2}$
688 358 979 309 308.3	6×10^{-16}	^{171}Yb$^+$, $6s\ ^2S_{1/2} \leftrightarrow 5d\ ^2D_{3/2}$
1064 721 609 899 145.3	1.9×10^{-15}	^{199}Hg$^+$, $5d^{10}6s\ ^2S_{1/2}$–$5d^96s^2\ ^2D_{5/2}$
1121 015 393 207 857.3	1.9×10^{-15}	^{27}Al$^+$, $3s^2\ ^1S_0 \leftrightarrow 3s3p\ ^3P_0$
1128 575 290 808 154.4	5×10^{-16}	^{199}Hg, $6s^2\ ^1S_0 \leftrightarrow 6s6p\ ^3P_0$

TABLE 10.5

Recommended Radiations of Stabilized Lasers

Frequency (THz)	Wavelength (fm)	Fractional Uncertainty	Absorber
1 267.402 452 899 92	236 540 853.549 75	3.6×10^{-13}	^{115}In$^+$, $5s^2\ ^1S_0 \leftrightarrow 5s5p\ ^3P_0$ transition
1 233.030 706 593 55	243 134 624.626 04	2.0×10^{-13}	^{1}H, 1S–2S two photon transition
1 064.721 609 899 143	281 568 867.591 969	1.9×10^{-14}	^{199}Hg$^+$, $5d^{10}6s^2\ ^2S_{1/2}\ (F = 0) \leftrightarrow 5d^96s^2\ ^2D_{5/2}\ (F = 2)\ \Delta m_F = 0$ transition
688.358 979 309 312	435 517 610.739 69	2.9×10^{-14}	^{171}Yb$^+$, $6s^2\ ^2S_{1/2}\ (F = 0, m_F = 0) \leftrightarrow 5d\ ^2D_{3/2}\ (F{=}2, m_F = 0)$ transition
642.121 496 772 6	466 878 090.061	4.0×10^{-12}	^{171}Yb$^+$,$^2S_{1/2}\ (F = 0, m_F = 0) \leftrightarrow {}^2F_{7/2}\ (F = 3, m_F = 0)$ transition
563.260 223 513	532 245 036.104	8.9×10^{-12}	^{127}I$_2$, a_{10} component, $R(56)$, 32-0 transition
473.612 353 604	632 991 212.58	2.1×10^{-11}	^{127}I$_2$, a_{16} component, $R(127)$, 11-5 transition
455.986 240 494 150	657 459 439.291 67	1.1×10^{-13}	^{40}Ca, $^1S_0 \leftrightarrow {}^3P_1$, $\Delta m_J = 0$ transition
444.779 044 095 5	674 025 590.8631	7.9×10^{-13}	^{88}Sr$^+$, $^2S_{1/2} \leftrightarrow {}^2D_{5/2}$ transition
385.285 142 375	778 105 421.23	1.3×10^{-11}	^{85}Rb, $5S_{1/2}\ (F_g = 3) \leftrightarrow 5D_{5/2}\ (F_e = 5)$ two photon transition
194.369 569 4	1 542 383 712	5.2×10^{-10}	^{13}C$_2$H$_2$, $P(16)(\nu_1 + \nu_3)$ transition
88.376 181 600 18	3 392 231 397.327	3×10^{-12}	CH$_4$, $F_2^{(2)}$ component, $P(7)\ \nu_3$ transition
29.054 057 446 579	10 318 436 884.460	1.4×10^{-13}	OsO$_4$

Following the recommendation of the CCTF (Consultative Committee for Time and Frequency) in 2017, there are now one microwave transition (hyperfine transition in ^{87}Rb) and eight optical transitions (both ion and neutral atoms) that are recommended as “secondary representations of the second”, as shown in Table 10.4. The frequencies of optical atomic clocks are measured against the primary Cs clocks directly or by means of international atomic time (TAI). The lowest uncertainties are 4×10^{-16}, only a factor of about two larger than that of the best primary Cs atomic clocks.

Based on the metre definition since 1983, length can be determined from the measurement of time interval. Meanwhile, the metre can also be practically realized by means of interferometric measurement with one of the radiations from a list provided by the CIPM. This list includes recommended radiations of stabilized lasers (Table 10.5) and values for radiations of spectral lamps and other sources [53]. Those frequencies and vacuum wavelengths can be used with uncertainties subject to the recommended conditions. For length metrology, the iodine-stabilized He-Ne laser operating at 633 nm is the most common because it is convenient to operate with an uncertainty of 2.1×10^{-11}.

Another important thing is that the index of refraction (n) of the light beam path must be included when transferring the length standard to practical length measurements. In air, the wavelength of the He-Ne laser at 633 nm is smaller than its vacuum wavelength by 2.7×10^{-4} for standard atmospheric conditions.

REFERENCES

1. Bureau International des Poids et Mesures (BIPM) Circular T, ftp://ftp2.bipm.org/pub/tai//Circular-T/cirthtm/.
2. Nicholson T. L., Campbell S. L., Hutson R. B., Marti G. E., Bloom B. J., McNally R. L., Zhang W., Barrett M. D., Safronova M. S., Strouse G. F., Tew W. L. and Ye J. 2015 *Nat. Commun.* **6** 6896.
3. McGrew W. F., Zhang X., Fasano R. J., Schäffer S. A., Beloy K., Nicolodi D., Brown R. C., Hinkley N., Milani G., Schioppo M., Yoon T. H. and Ludlow A. D. 2018 *Nature* **564** 87.
4. Recommendation CCTF 1 (2001) 2001 Comité Consultatif du Temps et des Fréquences, 15th Meeting, p. 132 (www.bipm.org/en/committees/cc/cctf/publications-cc.html#pv).

5. Matei D. G., Legero T., Häfner S., Grebing C., Weyrich R., Zhang W., Sonderhouse L., Robinson J. M., Ye J., Riehle F. and Sterr U. 2017 *Phys. Rev. Lett.* **118** 263202.
6. Riehle F., Gill P., Arias F. and Robertsson L. 2018 *Metrologia* **55** 188.
7. Hall J. L. 2006 *Rev. Mod. Phys.* **78** 1279.
8. Hänsch T. W. 2006 *Rev. Mod. Phys.* **78** 1297.
9. Yao Y., Jiang Y., Yu H., Bi Z. and Ma L. 2016 *Natl. Sci. Rev.* **3** 463.
10. Johnson L. A. M., Gill P. and Margolis H. S. 2015 *Metrologia* **52** 62.
11. Ma L., Jungner P., Ye J. and Hall J. L. 1994 *Opt. Lett.* **19** 1777.
12. Takano T., Takamoto M., Ushijima I., Ohmae N., Akatsuka T., Yamaguchi A., Kuroishi Y., Munukane H. and Miyahara B. 2016 *Nat. Photonics* **10** 662.
13. Rosenband T., Hume D. B., Schmidt P. O., Chou C. W., Brusch A., Lorini L., Oskay W. H., Drullinger R. E., Fortier T. M., Stalnaker J. E., Diddams S. A., Swann W. C., Newbury N. R., Itano W. M., Wineland D. J. and Bergquist J. C. 2008 *Science* **319** 1808.
14. Margolis H. S., Barwood G. P., Huang G., Klein H. A., Lea S. N., Szymaniec K. and Gill P. 2004 *Science* **306** 1355.
15. Madej A. A., Dubé P., Zhou Z., Bernard J. E. and Gertsvolf M. 2012 *Phys. Rev. Lett.* **109** 203002.
16. Huntemann N., Sanner C., Lipphardt B., Tamm Chr and Peik E. 2016 *Phys. Rev. Lett.* **116** 063001.
17. Godun R. M., Nisbet-Jones P. B. R., Jones J. M., King S. A., Johnson L. A. M., Margolis H. S., Szymaniec K., Lea S. N., Bongs K. and Gill P. 2014 *Phys. Rev. Lett.* **113** 210801.
18. S. M. Brewer, J-S. Chen, A. M. Hankin, E. R. Clements, C. W. Chou, D. J. Wineland, D. B. Hume, D. R. Leibrandt 2019 *Phys. Rev. Lett.* **123**, 033201.
19. Huang Y., Guan H., Liu P., Bian W., Ma L., Liang K., Li T. and Gao K. 2016 *Phys. Rev. Lett.* **116** 013001.
20. Chwalla M., Benhelm J., Kim K., Kirchmair G., Monz T., Riebe M., Schindler P., Villar A. S., Hansel W., Roos C. F., Blatt R., Abgrall M., Santarelli G., Rovera G. D. and Laurent Ph. 2009 *Phys. Rev. Lett.* **102** 023002.
21. Matsubara K., Hachisu H., Li Y., Nagano S., Locke C., Nogami A., Kajita M., Hayasaka K., Ido T. and Hosokawa M. 2012 *Opt. Express* **20** 22034.
22. Yao Huang, Huaqing Zhang, Baolin Zhang, Yanmei Hao, Hua Guan, Mengyan Zeng, Qunfeng Chen, Yige Lin, Yuzhuo Wang, Shiying Cao, Kun Liang, Fang Fang, Zhanjun Fang, Tianchu Li, Kelin Gao, 2020, *Phys. Rev. A* **102**, 050802.
23. Chou C. W., Hume D. B., Rosenband T. and Wineland D. J. 2010 *Science* **329** 1630.
24. Paul W., Osberghaus O. and Fischer E. 1958 *Forschungsberichte des Wirtschaft and Verkehrministeriums Nordrhein Westfalen* **415** 1.
25. Dehmelt H. G. 1968 Radiofrequency spectroscopy of stored ions I: Storage, advances in *Atomic and Molcular Physics* **3** 53.
26. R. Lange, N. Huntemann, J. M. Rahm, C. Sanner, H. Shao, B. Lipphardt, Chr. Tamm, S. Weyers, and E. Peik, 2020, *Phys. Rev. Lett.* Accepted
27. Yudin V. I., Taichenachev A. V., Oates C. W., Barber Z. W., Lemke N. D., Ludlow A. D., Sterr U., Lisdat Ch. and Riehle F. 2010 *Phys. Rev. A* **82** 011804.
28. Schmidt P. O., Rosenband T., Langer C., Itano W. M., Bergquist J. C. and Wineland D. J. 2005 *Science* **309** 749.
29. Dubé P., Madej A. A., Tibbo M. and Bernard J. E. 2014 *Phys. Rev. Lett.* **112** 173002.
30. Huang Y, Guan H, Zeng M, Tang L and Gao K 2019 *Phys. Rev. A* **99** 011401.
31. McFerran J. J., Yi L., Mejri S., Manno S. Di, Zhang W., Guéna J., Le Coq Y. and Bize S. 2012 *Phys. Rev. Lett.* **108** 183004.
32. Dicke R. H. 1953 *Phys. Rev.* **89** 472.
33. Katori H., Takamoto M., Pal'chikov V. G. and Ovsiannikov V. D. 2003 *Phys. Rev. Lett.* **91** 173005.
34. Dehmelt H. G. 1982 *IEEE Trans. Instrum. Meas.* **IM-31** 83.
35. Lemke N. D., Ludlow A. D., Barber Z. W., Fortier T. M., Diddams S. A., Jiang Y., Jefferts S. R., Heavner T. P., Parker T. E. and Oates C. W. 2009 *Phys. Rev. Lett.* **103** 063001.
36. Ludlow A. D. 2008 The strontium optical lattice clock: optical spectroscopy with sub-hertz accuracy (*PhD Thesis*), p. 153.
37. Campbell G. K., Boyd M. M., Thomsen J. W., Martin M. J., Blatt S., Swallows M. D., Nicholson T. L., Fortier T., Oates C. W., Diddams S. A., Lemke N. D., Naidon P., Julienne P., Ye J. and Ludlow A. D. 2009 *Science* **324** 360.
38. Ludlow A. D., Lemke N. D., Sherman J. A., Oates C. W., Quéner G., von Stecher J. and Rey A. M. 2011 *Phys. Rev. A* **84** 052724.
39. Porsev S. G. and Derevianko A. 2006 *Phys. Rev. A* **74** 020502.
40. Sherman J. A., Lemke N. D., Hinkley N., Pizzocaro M., Fox R. W., Ludlow A. D. and Oates C. W. 2012 *Phys. Rev. Lett.* **108** 153002.
41. Middelmann T., Falke S., Lisdat C. and Sterr U. 2012 *Phys. Rev. Lett.* **109** 263004.
42. Beloy K., Sherman J. A., Lemke N. D., Hinkley N., Oates C. W. and Ludlow A. D. 2012 *Phys. Rev. A* **86** 051404(R).
43. Yamanaka K., Ohmae N., Ushijima I., Takamoto M. and Katori H. 2015 *Phys. Rev. Lett.* **114** 230801.
44. Boyd M. M., Zelevinsky T., Ludlow A. D., Blatt S., Zanon-Willette T., Foreman S. M. and Ye J. 2007 *Phys. Rev. A* **76** 022510.
45. Lemke N. D. 2012 Optical lattice clock with spin-1/2 ytterbium atoms (*PhD thesis*) p. 44.
46. Ludlow A. D., Zelevinsky T., Campbell G. K., Blatt S., Boyd M. M., de Miranda M. H. G., Martin M. J., Thomsen J. W., Foreman S. M., Ye J., Fortier T. M., Stalnaker J. E., Diddams S. A., Le Coq Y., Barber Z. W., Poli N., Lemke N. D., Beck K. M. and Oates C. W. 2008 *Science* **319** 1805
47. Ludlow A. D., Boyd M. M., Ye J., Peik E. and Schmidt P. O. 2015 *Rev. Mod. Phys.* **87** 637.
48. Schnatz H., Lipphardt B., Helmcke J., Riehle F. and Zinner G. 1996 *Phys. Rev. Lett.* **76** 18.
49. Udem Th., Reichert J., Holzwarth R. and Hänsch T. W. 1999 *Phys. Rev. Lett.* **82** 3568.
50. Diddams S. A., Jones D. J., Ye J., Cundiff S. T., Hall J. L., Ranka J. K., Windeler R. S., Holzwarth R., Udem Th. and Hänsch T. W. 2000 *Phys. Rev. Lett.* **84** 5102.
51. Jones D. J., Diddams S. A., Ranka J. K., Stentz A., Windeler R. S., Hall J. L. and Cundiff S. T. 2002 *Science* **288** 635.
52. Telle H. R., Lipphardt B. and Stenger J. 2002 *Appl. Phys. B* **74** 1.
53. Quinn T. J. 2003 *Metrologia* **40** 103.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

11

Gravitation Measurements with Laser Interferometry

Harold V. Parks

CONTENTS

11.1 Introduction 171
11.2 Gravitational Wave Detection 172
    11.2.1 First Detection 172
    11.2.2 Detector Design 172
        11.2.2.1 Laser Source and Input Optics 172
        11.2.2.2 Fabry-Perot Cavities 173
        11.2.2.3 Interferometer Control and Gravitational Wave Signal Extraction 174
    11.2.3 Future Work on Earth and in Space 174
11.3 Absolute Gravimetry 174
    11.3.1 Falling Corner Cube Gravimeters 174
        11.3.1.1 Laser Source 175
        11.3.1.2 Falling Corner Cube 175
        11.3.1.3 Drag-Free Chamber and Drop Mechanism 175
        11.3.1.4 Super-spring Inertial Reference 175
        11.3.1.5 Signal Acquisition and Analysis 175
        11.3.1.6 Current Status and Future Work 175
    11.3.2 Cold Atom Gravimeters 176
        11.3.2.1 Atomic Fountain 176
        11.3.2.2 Atom Interferometry 176
        11.3.2.3 Raman Transitions and Atom Interferometery 177
        11.3.2.4 Inertial Reference 177
        11.3.2.5 Current Status of Cold-Atom Gravimeters 178
        11.3.2.6 New Methods 178
11.4 Determination of the Newtonian Constant of Gravitation 178
    11.4.1 Free-Falling Corner Cube Determination 179
    11.4.2 Atom Interferometer Determinations 179
    11.4.3 Suspended Fabry-Perot Cavity Determination 179
References 180

11.1 Introduction

Laser interferometry is an important tool for the study of fundamental physics. We are now in possession of atomic clocks with frequency accuracies of a few parts in 10^{16}. In recent decades great progress has been made in stabilizing lasers to these extraordinary time bases (and in fact laser technology is also key to the high accuracy of the most modern clocks). Since 1983, this confluence between high-accuracy time and optical measurements has been recognized in the SI system of units with the meter defined as the distance that light travels in a vacuum in exactly 1/299 792 458 s. Thus laser interferometry can be used not just for very precise measurements but also extremely accurate measurements directly traceable to the SI.

Here we describe the application of this tool to the measurement of gravitational effects. In particular to the detection of gravitational waves, the measurement of the local gravity acceleration with both macroscopic and atomic free-falling masses, and the measurement of the Newtonian constant of gravitation. Gravitational wave detectors take advantage of the sensitivity of laser interferometry to measure the extremely small displacements of a test mass in response to ripples in spacetime. Although in this application absolute accuracy is not critical, the laser stabilization techniques used for absolute interferometry are important to the operation of the instrument. Absolute accuracy is important for falling corner cube gravimeters, which can measure the local acceleration of gravity to a few parts per billion. Atom interferometer gravimeters are sometimes called matter wave interferometers. However they really measure the evolution of an atomic wave-function phase against a very well-defined laser field. Determinations of the gravitational constant require the measurement of the very small gravitational attraction from a well-characterized source mass and thus can benefit from both the sensitivity and the accuracy of laser interferometry.

11.2 Gravitational Wave Detection

11.2.1 First Detection

On 14 September 2015 the two Laser Interferometer Gravitational Wave Observatory (LIGO) detectors in Hanford, Washington and Livingston, Louisiana made the first direct detection of gravitational waves, recording the signal from two massive black holes which spiralled towards each other and coalesced roughly a billion light years from Earth (Abbot et al. 2016). The acceleration of the two black holes, each with a mass of approximately 30 suns, generated fluctuations in the curvature of spacetime that propagated outwards at the speed of light as a wave. The signal from the September 2015 event, denoted GW150914, is shown in Figure 11.1. The detector recorded displacements on the order of 10^{-18} m over the 4 km LIGO arm length. Between September 2015 and August 2017 it observed 5 black hole mergers and one neutron star merger which created a black hole. One of these black hole events was also detected by the Virgo detector in Italy (Abbot et al. 2017).

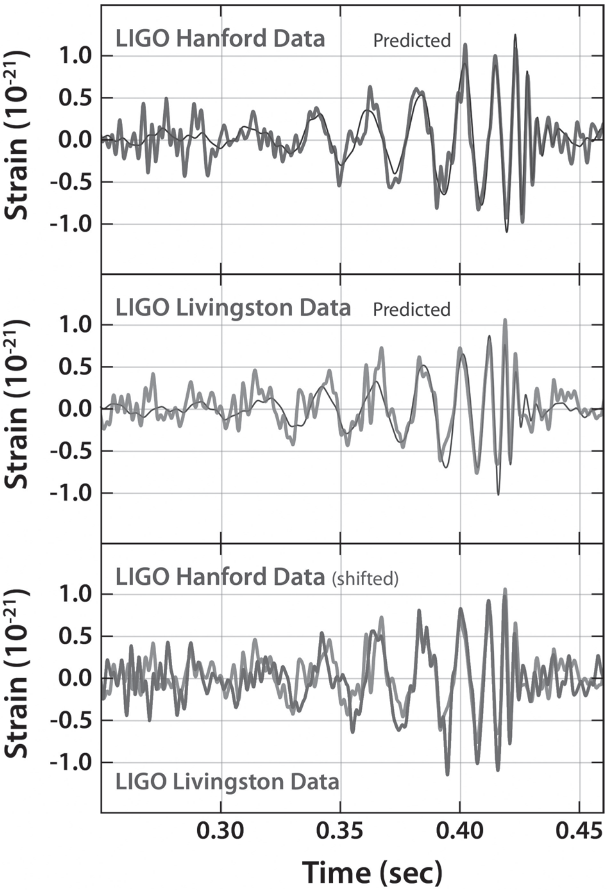


FIGURE 11.1 This is the signal representing the first direct detection of gravitational waves by the LIGO observatory. The gravitational waves shown here were generated when two black holes spiralled-in towards one another and merged. The signals from both LIGO detectors, one in Hanford, Washington and the other in Livingston, Louisiana are shown in the top two panels. The signal from each detector is shown along with a fit based on a model of two back holes spiralling-in towards each other and coalescing. The gravitational wave strain on the horizontal axes is the relative change in length of the two arms of the interferometer divided by the length of one arm. In the bottom panel, the signals are overlaid with a shift to account for the light delay between the detectors. The fact that the signals are from the two widely spaced detectors rules out local noise as a source of the signal. Image credit Caltech/MIT/LIGO Lab.

11.2.2 Detector Design

The LIGO instrument is part of a larger collaboration which encompasses the Virgo detector in Pisa, Italy (Acernese et al. 2014), and the GEO detector in Sarstedt, Germany (Dooley et al. 2016). The instrument shown in Figure 11.2 is the Advanced LIGO detector, though all of these instruments work on very similar principles. Advanced LIGO refers to the LIGO observatory after upgrades, completed in 2015, to the original two LIGO detectors (Martynov et al. 2016).

Because gravity exists with only one polarity, the lowest order propagating wave is a quadrupole mode, with space periodically expanding and the contracting along one axis while contracting then expanding (180° out-of-phase) along a perpendicular axis. The LIGO detector measures this gravitational wave strain with a Michelson interferometer with two perpendicular arms 4 km long. A Michelson interferometer divides a laser beam and sends it down two different paths with a beam splitter. Both beams are reflected back to the beam splitter and recombined. Interference between the two beams converts changes in the relative path lengths of the two interferometer arms to changes in the intensity of the light output from the interferometer. The LIGO Michelson interferometer is modified by Fabry-Perot mirrors that reflect light multiple times back and forth the length of each arm, thus effectively multiplying the interferometer arm length. The instrument is designed to be sensitive to gravitational waves with frequencies in the range of 10 Hz to 10 kHz. The low end of this range is set by seismic and thermal vibrational noise that masks the gravity wave signal. At the high end the effective arm length limits the sensitivity. As an antenna, the detector is most sensitive to signals with a wavelength equal to four times the effective arm length. For higher frequency signals the time delays due to the light transit time through the arms begin to average down the signal. In addition laser beam intensity fluctuations due to quantum shot noise increase with frequency.

11.2.2.1 Laser Source and Input Optics

The laser source is a Nd:YAG laser which produces 1064 nm light. This laser is capable of producing a beam with nearly 200 W of power but has been held near 20 W for the initial observing runs. As the photon pressure on the mirrors is much greater than the forces expected from gravitational waves, the laser must be extremely stable and multiple feedback loops

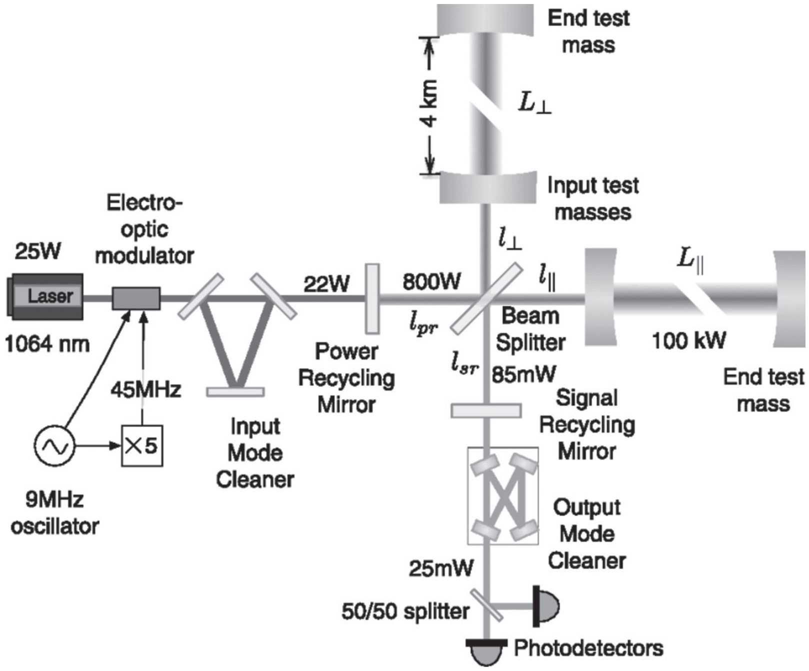


FIGURE 11.2 A simplified schematic of a LIGO detector is shown here. The modified Michelson interferometer is fed with a laser that produces 1064 nm light. The beam goes through an electro-optic modular that impresses a RF phase modulation on the beam. This modulation is used to generate error signals that control servos that position the interferometer optical elements. Mode cleaners at the interferometer input and output clean up the laser beam shape. Two mirrors in each interferometer arm form a Fabry-Perot cavity resonance that allows power to build up in the cavity and increase the detector signal to noise. These mirrors are suspended as pendulums and it is the displacement of these mirrors that respond to the gravitational waves. The power-recycling and signal-recycling mirrors further increase the signal to noise. The Michelson interferometer is held near a dark fringe so the output power is very low. The laser is designed to produce nearly 200 W, but on the first observing run, the laser power was held to 25 W. Image credit (Martynov et al. 2016).

are used to reduce the intensity noise. After the beam leaves the laser it passes through an electro-optic modulator (see Figure 11.2). This electro-optic modulator impresses a radio frequency (RF) modulation on the laser beam which is used by the system that controls servos to adjust the mirror and beam splitter positions and alignments. The beam also passes through a telescope for matching the beam size output from the laser to the beam size used in the interferometer, a Faraday isolator to prevent reflected light from feeding back to the laser and mode cleaner to clean up the beam shape. The Faraday isolator and all subsequent optics and beam paths, including the entire length of the 4 km arms, are in ultra-high vacuum. Optical baffles are carefully placed throughout the interferometer so that stray light, which may be Doppler-shifted after reflecting from a vibrating surface, is eliminated.

11.2.2.2 Fabry-Perot Cavities

The main Fabry-Perot cavities in the two arms are each formed by two mirrors facing one another. One mirror is only partially (98.6%) reflective so some of the laser beams can pass through it. Once inside a photon will, on average, bounce back and forth between the mirrors a number of times before exiting back out the partially reflective mirror. The laser beam in the arm cavities has a diameter of about 5 cm. This large beam size reduces the effect of thermal vibration noise in the mirrors as it averages over a large area. These mirrors are suspended pendulum-like from steel wires and are the test masses that move in response to gravitational waves. They are fused silica cylinders with a mass of 40 kg and a diameter of about 34 cm and length of 20 cm. The mirror is hung from a suspension which has multiple passive and active vibration isolation stages.

The Michelson interferometer output is held near a dark fringe which means that most of the light input to the interferometer is reflected back towards the laser. This reduces the sensitivity of the Michelson output signal to laser intensity fluctuations. A partially reflecting mirror is put in the path of the beam between the laser and the beam splitter to recycle this light. This creates an optical resonance that multiplies the power inside the cavity by a factor of 38 with respect to the power input to the cavity (this multiplication factor is set by the reflectivity of the mirror which is 97%). The arm cavities multiply the power by an additional factor of 270, so when the laser is operating at its full designed power of nearly 200 W, the power in each arm cavity reaches close to 1 MW. Another partially reflective mirror at the output port of the interferometer enhances the signal.

The optics will absorb a small amount of this light, causing distortions due to heating. These distortions in the optics are probed with 532 nm lasers and super luminescent LEDs. Heaters around the diameter of the arm cavity mirrors counteract the measured distortions. The beams also pass through a compensation plate immediately before they enter each arm cavity. CO_2 laser projectors are fired at the surface of the two compensation plates warping the surface of the plates in a pattern that cancels the distortions of the cavity mirrors.

11.2.2.3 Interferometer Control and Gravitational Wave Signal Extraction

The interferometer has four Fabry-Perot cavities: the two arm cavities, a cavity formed by the power recycling mirror and a cavity formed by the signal recycling mirror. All of these cavities must be held in resonance, with an integer number of laser beam wavelengths in one round trip through each cavity. Most of the interferometer optical elements are suspended for vibration isolation, and magnet and coil actuators allow the power and signal recycling mirrors as well as the beam splitter to be moved to keep the cavities on resonance and the interferometer output near a dark fringe. The arm mirrors are moved with a combination magnet and coil (on the top three stages of the quadruple pendulum) and electrostatic actuators (on the mirrors themselves). However, instead of acting against a fixed surface, the actuators on the arm mirrors act against another pendulum. This is to prevent the actuators from generating vibrations in the pendulum-supporting structures that could upset the mirrors.

The Pound–Drever–Hall method is used to generate the servo error signal that is fed to the actuators. An electro-optic modulator impressed an RF phase modulation on the laser beam. Light reflected back from a Fabry-Perot cavity contains a mixture of light coming back from inside the cavity as well as light reflected directly back from the partially reflecting mirror. This mixing converts the phase modulation to an intensity modulation that encodes the frequency difference between the laser light and the cavity resonance. In the LIGO interferometer light reflected from various cavities is picked off and used to generate error signals that are fed to the power and signal recycling mirror-positioning servos. Common-mode length changes of the arm cavities are fed back to the laser controller so that the laser frequency is locked to the average arm cavity length. Quadrant photo-detectors also use the RF signal for alignment of the various optics. The output of the Michelson interferometer is passed through a mode cleaner, which removes light from higher order transverse beam modes as well as the RF intensity modulation. The output light is then detected and fed back to control the differential length of the arm cavities. Of course, differential length changes can be due to gravitational waves, and the gravitational wave signal is derived from a combination of the photo-detector output and the differential arm length servo signal. This signal is calibrated against the photon pressure when a 2 W laser is reflected from each of the arm cavity end mirrors.

11.2.3 Future Work on Earth and in Space

LIGO continues to be upgraded. In the near future, improvements to the optics are planned and squeezed light will be injected into the interferometer output to reduce phase noise (Aasi et al. 2013). A third LIGO detector is planned for India. And the Kamioka Gravitational Wave Detector, currently under construction in Japan, will use cryogenically cooled optics to reduce thermal vibrations (Aso et al. 2013).

A space-based detector, the Laser Interferometer Space Antenna (LISA), is in the planning stages by the European Space Agency for launch in the 2030s (Amaro-Seoane et al. 2017). This detector will be sensitive to much lower frequency signals, 0.1 mHz to 100 mHz, than ground-based detectors. Three identical satellites in a heliocentric orbit will be arranged in a triangle 2.5×10^6 km on a side. Each satellite will aim a several watt 1064 nm laser at the other two satellites. The distances between the satellites are so great that passive return reflections cannot be utilized for interferometry due to beam divergence. So the lasers on each satellite are phase-locked to the incoming beams. The optical systems will be floating drag-free inside the satellites so that they are shielded from external forces such as solar radiation. In 2015 the LISA Pathfinder mission was launched (Armano et al. 2018). This was a satellite in Earth orbit containing two optical platforms making up a short-interferometer arm. This was much too small to detect gravitational waves but it did successfully test the technology that will be used in the full LISA mission.

11.3 Absolute Gravimetry

Measurements of the gravity field at the surface of the Earth provide information about the interior structure of the earth as well as variations in local geology. An accurate measurement of the acceleration of gravity g is also important for the redefinition of the kilogram. The weight of a test mass is balanced against a magnetic force that can be expressed in terms of the Plank constant and the electron charge. To convert this force measurement to mass, g must be known at the point of weighing.

Spring-type gravimeters, which measure the force of gravity on a cantilevered arm supported by a spring, and superconducting gravimeters, which replace the cantilevered arm with a superconductor levitating in a magnetic field, are very sensitive to changes in gravity. But they must be calibrated and their sensitivity drifts over time. The most accurate way to measure gravity is to drop a test mass and observe its acceleration. Instruments that drop masses with a reflector attached and instruments that drop laser-cooled atoms have both reached accuracies approaching one part in a billion for g at the surface of the Earth.

11.3.1 Falling Corner Cube Gravimeters

A simplified schematic of the FG5 gravimeter is shown in Figure 11.3 (Niebauer et al. 1995). This gravimeter was developed at JILA and is now commercially available. Most other (macroscopic object) free-fall gravimeters use a similar design. A modified Mach-Zehnder interferometer is used to track the motion of a falling test mass in vacuum. A Mach-Zehnder interferometer works like a Michelson interferometer except that the split beams are recombined with a second beam splitter rather than being reflected back to the first beam splitter.

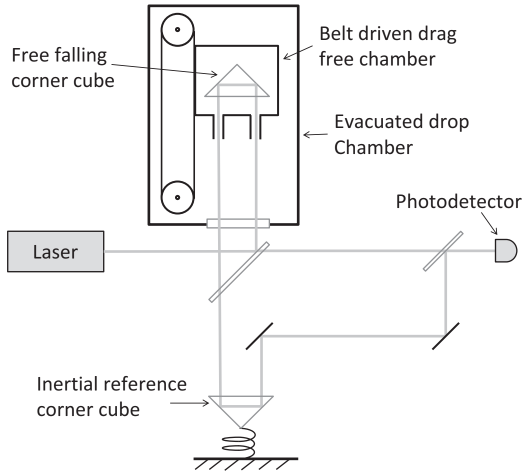


FIGURE 11.3 This is a simplified schematic of the FG5 free-falling corner cube gravimeter. A Mach-Zehnder interferometer monitors the position of a corner cube falling in vacuum with respect to a second corner cube that is carefully isolated from vibration. A drag-free chamber follows the falling corner cube and sweeps away residual gas. The drag-free chamber also gently catches the falling corner cube at the bottom of the drop and raises it for the next measurement.

11.3.1.1 Laser Source

For best accuracy an iodine-stabilized helium-neon (HeNe) laser is used. This type of laser is frequency-stabilized to an absorption in an iodine vapour cell. The frequency of this absorption line is well characterized so the laser wavelength can be traced to the SI. A calibrated polarization-stabilized He-Ne laser or Zeeman-stabilized laser may also be used. These lasers are slightly less accurate but are more inexpensive and robust. The laser is typically delivered to the interferometer with a fibre optic. The iodine-stabilized laser requires a Faraday isolator to prevent the return laser light from interfering with the lock to the iodine absorption line. For polarization and Zeeman-stabilized lasers, a simple quarter wave plate and polarizing beam splitter can be used to prevent back-reflections.

11.3.1.2 Falling Corner Cube

The laser beam reflects from a corner cube prism embedded in the falling test mass. A corner cube is made up of three perpendicular intersecting flat planes that form a retroreflector. The corner cube has the useful property that the path length of the laser beam is insensitive to translations of the corner cube perpendicular to the beam as well as by rotations around the optical centre of the corner cube. Thus, the falling mass is designed to have its centre of mass coincide with the optical centre of the corner cube.

11.3.1.3 Drag-Free Chamber and Drop Mechanism

The falling mass is enclosed in a drag-free chamber. This chamber is connected to a motorized drive and is servoed so that it follows the mass and prevents drag due to residual gas in the chamber. Tubes which allow the laser beam to enter and leave the chamber are as effective as a window without the wedge errors associated with such a moving optical element. The drag-free chamber also gently catches the falling mass at the bottom of the drop and then raises the mass. The mass is released for the next drop when the chamber is briefly driven down with an acceleration of slightly more than 1 *g*.

An alternate strategy for dealing with drag is the rise and fall method employed by an instrument built at the Istituto Nazionale di Ricerca Metrologica in Torino, Italy (D'Agostino et al. 2008). In this scheme, the mass is tossed upwards and measurements are taken as the mass rises then falls back to where it started. Since the sign of the drag is opposite during the rising and falling parts of the trajectory, the effect of drag cancels when the acceleration is averaged over the full trajectory.

11.3.1.4 Super-spring Inertial Reference

A second corner cube is attached to an inertial reference. This reference mass is well isolated from any vibration using a "super-spring". This is a two stage isolation scheme, with a servo-controlled active stage supporting a long period spring from which the corner cube hangs. The effective period of this super-spring is more than 60 s. An important feature of this design is that the reference corner cube is in line with the free-falling corner cube so that ground tilts in response to the release and catch of the falling mass do not couple to the interferometer signal through the reference mass. The evacuated drop chamber also sits on a separate tripod from the one supporting the rest of the interferometer (which is not in vacuum) to reduce vibration coupling to the reference mass. Vibrations of the drop chamber can still affect the signal through air-gap modulation when the drop chamber window moves causing the relative laser path lengths in vacuum and air to change. A pool of alcohol can be placed on top of the super-spring during set-up of the instrument so that the reflection off the surface of the pool can be used to adjust the verticality of the laser beam.

11.3.1.5 Signal Acquisition and Analysis

The original FG5 design has a drop length for the falling mass of approximately 20 cm. So the interference fringe signal, recorded by an avalanche photodiode, sweeps from DC as the mass starts out at rest to 6 MHz in 0.2 s. A time-interval counter referenced to a rubidium-stabilized clock records the time of a subset of the zero-crossings recorded by the detector. A least-squares fit to the recorded data is used to find the acceleration of gravity *g*. This fit must take into account the vertical gravity gradient and use a retarded time to account for transit time of the light between the falling mass and the photodetector.

11.3.1.6 Current Status and Future Work

The major sources of uncertainty in FG5 measurement include delays in the electronics, air gap modulation, rotation of the dropped corner cube, Coriolis effects due to the rotation of the earth which cause a horizontal translation of the falling mass, and imperfections in the optics which affect the parallelism of the test and reference beams. This accuracy of this type of instrument is much smaller than the actual variation of gravity due to tides, atmospheric density, polar motions of the earth, and the water table. To a large extent these effects can be modelled and removed from the signal, though the water table gravity can be particularly difficult. However, because of these unavoidable sources of environmental noise, much recent work has been focused on making the instruments more portable and easier to use (Vitushkin and Faller 2002; Hu et al. 2010).

11.3.2 Cold Atom Gravimeters

An alternative type of absolute gravimeter uses laser-cooled atoms instead of a macroscopic free-falling mass. This is an active field of research and a number of different atomic interferometers have been constructed, but Figure 11.4 shows a simplified schematic of a pioneering experiment at Stanford (Peters, Chung and Chu 1999 and 2001). The features shown in Figure 11.4 are still common to most cold-atom gravimeters.

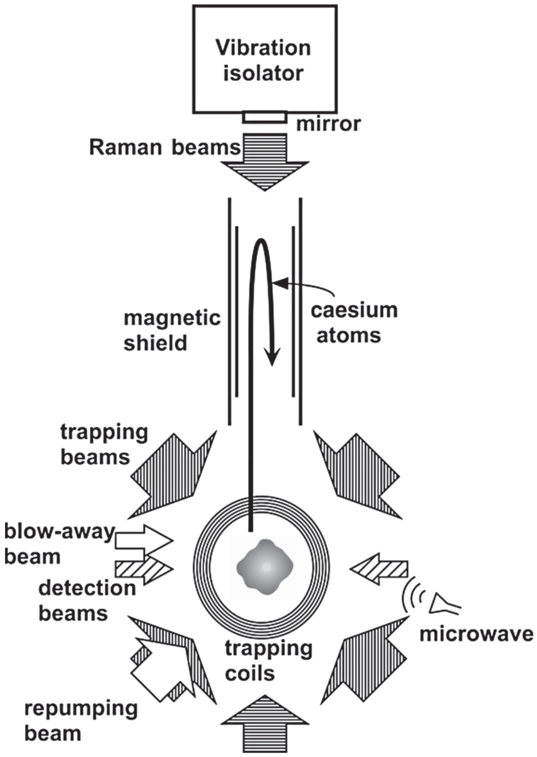


FIGURE 11.4 A simplified schematic of a cold atom gravimeter constructed at Stanford is shown (Peters, Chung and Chu 1999 and 2001). A MOT is used to trap and cool a cloud of caesium atoms. These atoms are then launched up into the magnetically shielded chamber. As the atoms are launched, microwaves and lasers prepare the atoms in one of the caesium ground-state hyperfine sub-levels. Two counter-propagating laser beams are tuned to drive stimulated Raman transitions between the two ground-state hyperfine sub-levels. A sequence of carefully controlled pulses from the Raman laser converts small changes in the laser Doppler shift as seen by the falling atoms into changes in the relative population in the two hyperfine sublevels. This signal is read out by observing fluorescence induced by the detection laser once the cloud of atoms falls back to the trap region allowing the gravitational acceleration to be calculated. Image credit (Peters, Chung and Chu 1999).

11.3.2.1 Atomic Fountain

An atomic fountain is used to launch a cloud of caesium or rubidium atoms. This technique is well established for use in atomic clocks. The atoms are first trapped in a magneto-optical trap (MOT). The MOT uses three pairs of counter-propagating laser beam to slow the thermal motion of a dilute gas of atoms and a magnetic field to confine them at the centre of the trap. The cloud is cooled to a temperature of a few μK and then released from the trap. In many gravimeter designs, the trapping magnetic field is turned off and the frequency of the trapping laser beams is ramped producing a moving optical molasses which launches the atoms upwards, though in some instruments the trap is simply turned off allowing the atoms to fall. The ground state of caesium and rubidium has two hyperfine sub-levels. The atoms are prepared in one of these sublevels using a microwave or laser source. A magnetic field is also briefly applied so atoms in the $m_F = 0$ magnetic sub-level can be selected. All atoms not in the correct state are blown away with a laser tuned to an atomic resonance. During this process, atoms with a narrow vertical velocity distribution can also be selected from the μK cloud, resulting in vertical temperature distribution as low as 10 nK. The result is a cloud on the order of 10^7 atoms approximately a millimetre across moving only under the influence of gravity (once the atoms enter a region that is shielded from magnetic fields).

11.3.2.2 Atom Interferometry

A series of laser pulses is used to measure the effect of gravity on the cold-atom cloud. To understand how this works, first consider a two-state atom that absorbs a photon that takes the atom from the lower to the upper energy level. The atom receives a kick when it absorbs the momentum of the photon and will receive another kick if it emits a photon and returns to the ground state. This photon recoil momentum is exploited to cool the atoms in the MOT, but here it is used for a different purpose. When an atom interacts with a coherent laser field it will undergo Rabi oscillations that transfer the state cyclically between the two states through absorption and stimulated emission. If the laser beam is turned off part way through a Rabi cycle the atom is left in a super-position of the two energy levels. However, this means that the atom is also in a superposition of two momentum states.

To measure gravity, a laser pulse takes the atoms in the cloud produced by the fountain from the initial state to a superposition of two states. The wave function of an atom localized in this cloud will be split into a super-position of two wave packets that are moving apart. This pulse is called a π/2 pulse. After a time delay a π pulse, twice as long as the first pulse, reverses the momentum of the wave packets sending them back towards one another. Finally, a second π/2 pulse is applied. This is an analogue to the Mach–Zehnder interferometer where the beam splitters are replaced with π/2 pulses and the mirrors are replaced with a π pulse. In the absence of gravity (and if the optical field is held perfectly resonant with the atomic transition), the atom will end up back in the initial state after the final π/2 pulse. However, a gravitational field is present and if the laser is aligned with the vertical axis, then one path is sent higher than the other and the wave function for that path will acquire an additional phase shift due to the gravitational potential energy. In this case the atoms end up in a superposition of states that depend on the value of *g*. The population in each state can be read out with laser fluorescence spectroscopy.

The interferometer phase difference due to the gravitational field, up to small corrections due to the gravity gradient, depends on the gravitational acceleration *g* as $\Delta\varphi = kgT^2$, where *k* is the wavenumber of the photon that the atom absorbs or emits when transition from one state to the other and *T* is the time interval between laser pulses (Peters, Chung and Chu 2001). The photon recoil momentum is proportional to *k*, and *T* is limited by the time the atom could remain in free-fall. Thus the sensitivity of the gravimeter is proportional to the photon momentum and to the square of the free-fall time.

Another way of looking at the atom interferometer is from the frame of reference of the freely falling atom. It sees the effect of the gravitational field as a Doppler shift which alters the phase of the optical field. This suggests a mode of operation that is used in most atom gravimeters: the optical frequency is chirped to cancel the Doppler shift of the falling atoms so that the phase shift between the two interferometer paths remains near zero. In either perspective, the key point is the way that atomic phase evolves with respect to the optical field. The chirp rate can be varied slightly from measurement to measurement to adjust the phase difference and scan the interferometer across the fringes. A scan from the Stanford gravimeter is shown in Figure 11.5. The value for *g* can be found by fitting this scan to a sine function and taking the phase of the fitted function. The fringe contrast is the ratio of the amplitude of the fringes to the mean value of the signal and for a perfect interferometer this would be 1. In Figure 11.5, the fringe contrast is somewhat less than 1 because the Raman lasers do not produce perfect π/2 and π pulses.

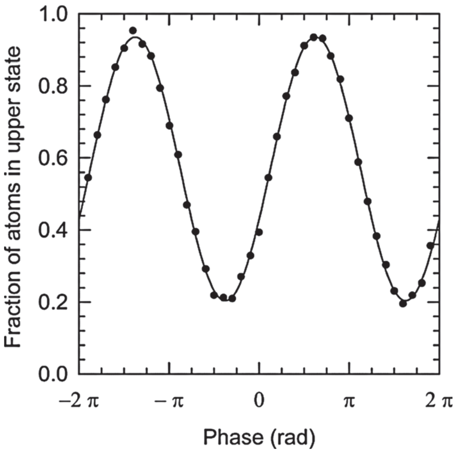


FIGURE 11.5 This is the signal from the interferometer shown in Figure 11.4. Each point represents a single-atom cloud launch with the Raman laser frequency shift adjusted after each point to scan across the interference fringe. This scan took 1 min to complete. Image credit (Peters, Chung and Chu 1999).

11.3.2.3 Raman Transitions and Atom Interferometery

Atom interferometer gravimeters have been demonstrated with single-photon transitions as described in the last section. But the requirement on the laser frequency stability is extremely high. The optical phase must remain coherent with the transition over the half second or so free-fall time in a typical gravimeter. This means that the frequency must be stable to a few Hz. At optical frequencies, this is a relative stability on the order of a part in 10^{14}. An easier method is to use two-photon Raman transitions. Two counter-propagating laser beams are used that have a frequency difference equal to the hyperfine splitting of the ground state. The atomic state is driven between the two hyperfine levels by absorbing a photon from one laser beam while simultaneously emitting a photon in the opposite direction into the other laser beam. The result is that the atom receives a kick with a magnitude equal to the momentum of two optical photons. The interferometer phase depends only on the frequency difference between the two lasers and not the absolute frequency of each laser, so the stability requirement is reduced to a few Hz out of 10 GHz. The sensitivity of the gravimeter is proportional to the recoil momentum, so the fact that the atom gains the momentum of two photons is an additional benefit.

In practice, external cavity diode lasers can provide the extremely narrow linewidth necessary for these experiments and the beat frequency between the two lasers can be locked to a frequency source with a very stable time base such as a rubidium clock. To create the counter-propagating laser beams, two beams with the same linear polarization are overlapped and sent through the apparatus in the same direction. A quarter wave plate in front of a mirror reflects the beam directly back along its path with the polarization rotated 90°. A similar scheme can be used with circularly polarized light. Transition rules only allow the Raman transition with photons with orthogonal polarizations. This leaves two pairs of counter-propagating beams that can drive the transition. But since the atoms are moving, the lasers can be tuned so that the Doppler shift puts the atoms in resonance with only one pair of beams.

11.3.2.4 Inertial Reference

The mirror that reflects the Raman beams defines the inertial reference that the gravitational acceleration is measured against. The Raman beams will remain in resonance with the Raman transition if they are both shifted in frequency by the same amount since the frequency difference is the critical parameter. The frame commoving with the mirror is the only frame where the Doppler shift of the incident beam and reflected beam is the same. As with the reference corner cube in the FG5, this mirror must be carefully protected from vibration generally with a combination of passive springs and an active feedback system. The system is much less sensitive to vibrations in the other optical components.

11.3.2.5 Current Status of Cold-Atom Gravimeters

Transportable atom gravimeters have a fountain height on the order of 10 cm allowing for a time between Raman pulses of slightly more than 0.1 s. If the Raman laser frequencies are near 800 nm, the acceleration of gravity near the surface of the Earth gives a phase shift of several million radians between the two paths. Atom gravimeters can approach an accuracy of 10^{-9} g and compare favourably to the FG5 gravimeter (Gillot et al. 2014). With a MOT and two phase-locked narrow linewidth Raman lasers, the optical system of an atom gravimeter is quite complex compared to the FG5. But the advantage of having no (macroscopic) moving parts should not be underestimated. Atom gravimeters have only recently reached a level of maturity where they can be used for field surveys while the FG5 has been used routinely for several decades, so it is not yet completely clear how the two technologies will compete in practical field use.

The atom gravimeter can be extended to a gravity gradiometer. Two atomic fountains are arranged, either side by side or one above the other, and the atoms in each fountain are addressed with the same Raman beams (McGuirk et al. 2002). Gradiometery is interesting because it is sensitive to local density inhomogeneities. Since the same Raman beam probes each atomic cloud, the system is much more tolerant of vibrations than a single gravimeter. Gradiometer-like designs with widely separated atomic fountains have also been proposed for use as gravitational wave detectors. Work has begun on a demonstration project with a several hundred metre laser cavity by the Matter-wave laser Interferometric Gravitation Antenna (MIGA) collaboration in France (Geiger et al. 2015).

11.3.2.6 New Methods

Recent work has been directed at using Bragg diffraction as an alternative to Raman diffraction in gravimeters (Mazzoni et al. 2015). The two counter-propagating laser beams form a grating that can diffract an atom without changing its internal state. This Bragg diffraction process can be thought of as similar to Raman diffraction, where the atom absorbs a photon from one beam and emits a photon into the other beam. Except with Bragg diffraction, the frequency difference between the two laser beams is tuned not to connect two different internal atomic energy levels, but two energy levels that differ only by the kinetic energy of the atom before and after it has received the photon recoil momentum kick. An advantage of the Bragg diffraction is that the frequency difference can be tuned so that the atom absorbs and emits not one pair of photons but multiple pairs gaining a much larger momentum kick, thus improving the sensitivity of the gravimeter. Raman transition interferometers are sensitive to Zeman and AC Stark shifts that act differently on the two electronic states. These effects are less of a problem with Bragg interferometers. In addition, Bragg diffraction allows interferometry with atoms such as calcium, strontium, and ytterbium that have transitions that are convenient for laser cooling but lack the hyperfine structure of caesium and rubidium. These atoms do not have an electronic magnetic moment in their ground state since they do not have an unpaired valance electron and some isotopes have no nuclear moment making them insensitive to external magnetic fields. Atom interferometry with Bragg diffraction is less well developed than gravimeters employing Raman diffraction, but a gravimeter using ^{88}Sr has been demonstrated in the laboratory with up to eight photon recoils.

A phenomenon related to Bragg diffraction is Bloch oscillation. This is a quantum mechanical effect where particles trapped in a periodic potential will undergo oscillations when a constant force is applied. These oscillations have been observed with atoms moving under the influence of gravity in a potential formed by an optical grating generated from a laser reflected from a mirror, and the oscillation frequency used to calculate the gravitational acceleration (Ferrari et al. 2006). Work has also been done on using Bose-Einstein condensates in gravimeters (Abend et al. 2016). This can pack the atoms into a cloud with a diameter 100 times smaller than the thermal clouds and thus reduce issues due to laser beam nonuniformity. A gravimeter has also been demonstrated using a single-photon optical clock transition in strontium. The laser requirements are very demanding for a single-photon interferometer since the optical phase itself must be controlled rather than a microwave phase difference (Hu et al. 2017). The optical clock transition is so named as it has an extremely narrow linewidth that is useful for optical clocks because the width of the spectral line is inversely proportional to the noise in the phase evolution when the atom is in a superposition of the ground and excited states. When the interferometer is used in a gradiometer mode, phase noise in the Raman beams cancels allowing the measurement to take advantage of the narrow line width of the clock transition. Test of the equivalence principle is also being conducted with atom interferometers that simultaneously measure the acceleration due to gravity of different atomic species.

11.4 Determination of the Newtonian Constant of Gravitation

The constant of gravitation *G* is the constant of proportionality in Newton's law of gravitation, setting the strength of the gravitational force. To measure this constant we must have a gravitational source mass and test mass both with well-characterized mass and density distribution. In practice this means that the masses must be carefully manufactured objects of uniform density and known size that are small enough to be placed on a mass balance. The tiny gravitational pull of such a mass must be measured in the presence of Earth's gravitational field. The result is that, even though the gravitation force is ubiquitous in everyday life, the accepted value of *G*, $(6.674\,08 \pm 0.000\,31) \times 10^{-11}$ m^3 kg^{-1} s^{-2}, has a very high uncertainty compared with other universal constants (Mohr, Taylor and Newell 2012).

From the classic measurement by Mitchell and Cavendish published in 1798 (Cavendish 1798) through the 2000 experiment at the University of Washington that obtained the most precise measurement to date (Gundlach and Merkowitz 2000), torsion balances are the most common apparatus used to determine *G*. A dumbbell-shaped test mass is hung from a thin wire or fibre. This produces an instrument that is very sensitive to horizontal gravity gradients but is insensitive to the gravitational field of the Earth. Laser interferometers are not particularly useful for measuring rotations and have not played a big role in this type of experiment. However, the material properties of the fibre that produce the balance's restoring torque are not necessarily straightforward. This has led to interest in alternative methods for determining *G*, some of which have employed laser interferometers.

11.4.1 Free-Falling Corner Cube Determination

An experiment at JILA used an FG5 gravimeter to determine *G* with an uncertainty of 0.14% (Schwarz et al. 1998). A 500 kg source mass consisting of tungsten billets arranged into a ring configuration fitting around an FG5 drop chamber was alternately moved above then below the gravimeter causing the local gravitational acceleration to change by a bit less than a part in 10^7. The ring configuration of the source mass not only fits around the vacuum chamber of the gravimeter as the source mass moved up and down but it also was designed to produce a saddle point in the gravitational field near the location of the falling mass. This reduced the sensitivity of the gravitational force to errors in the relative position between the source and test masses. The accuracy of this measurement was limited by the fact that it was measuring the signal tiny acceleration of the source masses on top of the acceleration from Earth's gravity. Replacing the lower stationary reference corner cube in the FG5 with a second free-falling mass would create a gradiometer configuration and cause the uniform gravitational field from Earth to cancel out of the signal leaving the signal from the source masses along with the small vertical gradient of Earth's field. A proposed variation of this experiment takes this a step further with a differential gradiometer configuration using three free-falling masses one above another (Rothleitner and Francis 2014). This cancels both the uniform and linear gradient of Earth's field leaving only the second derivative of the gravitational field which will be dominated by the 500 kg source mass. The mass positions in the JILA experiment had to be moved from one position to the other every 20 min to cancel the drift in the local gravity due to tides. The movement of this large amount of mass can cause errors through ground tilts and other effects. The differential gradiometer method reduces the need for this mass modulation.

11.4.2 Atom Interferometer Determinations

Proof of principle experiments to measure *G* using both vertical (Fixler et al. 2007) and horizontal (Biedermann et al. 2015) Raman atom interferometers have been carried out at Stanford. The horizontal interferometer was also used to demonstrate a method to test the inverse-square law for gravity. The Misura Accurata di Gmediante Interferometria Atomica (MAGIA) experiment at Università di Firenze has measured *G* to an accuracy of 0.015% (Lamporesi et al. 2008; Rosi et al. 2014). The MAGIA instrument is a Raman gradiometer that uses a single atomic fountain to launch two clouds of rubidium atoms, one after another, with different velocities such that they simultaneously reach an apogee of 60 cm and 90 cm above the MOT (Figure 11.6). A 500 kg ring of tungsten billets was arranged in a ring around the waist of the interferometer and positioned between the measurement regions for each cloud. The ring was spilt horizontally into two halves and one half moved above the top cloud and the other below the bottom cloud, reversing the sign of the gravitational gradient generated by the source mass. This cycle was repeated about once an hour. This apparatus has also been used to produce three simultaneous rubidium clouds to demonstrate a proof of principle differential gradiometer measurement of *G* like that proposed for corner cube gravimeters (Rosi et al. 2015).

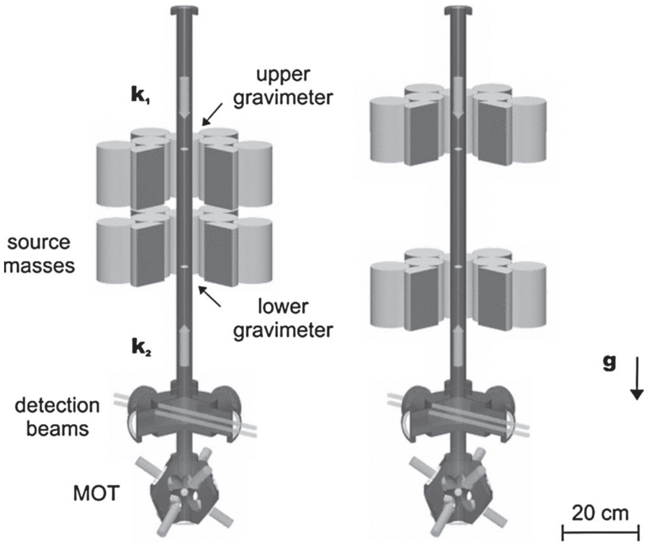


FIGURE 11.6 Here the MAGIA apparatus for determining the Newtonian constant of gravitation is shown (Lamporesi et al. 2008l; Rosi et al. 2014). A cold-atom gradiometer is implemented using a single MOT to launch two clouds of atoms so they reach apogee 30 cm apart at the locations marked as upper and lower gravimeter in the diagram. Tungsten billets with a total mass of 500 kg are arranged in a ring around the apparatus. When they are moved from the position shown on the left to the position shown on the right, the gravity gradient due to these source masses is reversed. Image credit (Lamporesi et al. 2008).

11.4.3 Suspended Fabry-Perot Cavity Determination

An experiment at JILA measured *G* using a Fabry-Perot interferometer to monitor the spacing between two suspended mirrors in a gradiometer configuration (Parks and Faller 2010). A He-Ne laser was locked to the 30 cm cavity formed by the mirror. When the 500 kg tungsten source masses moved, the gravitational pull on the mirrors caused the cavity length to change by 90 nm. The signal was read out by beating the laser with another laser locked to a cavity in the pendulum support so that length changes due to any thermal expansion and contraction of the supports cancel out. Unlike the free-fall experiments, the mirror displacement signal has to be calibrated to find the force, and this was done measuring the period of oscillation when the mirrors were allowed to swing freely. Although this experiment claimed an uncertainty well below 0.01%, it found a value that differs from the accepted value of *G* by 0.026% (Mohr, Taylor and Newell 2012). However, *G* measurements are very difficult and there are likely many unexpected and unknown sources of errors.

REFERENCES

Aasi, J., Abadie, J. and Abbott, B. P. et al. 2013. Enhanced sensitivity of the LIGO gravitational wave detector by using squeezed states of light. *Nature Photonics* 7: 613–619.

Abbot, B. P., Abbott, R. and Abbott, T. D. et al. 2016. Observation of gravitational waves from a binary black hole merger. *Physical Review Letters* 116: 061102.

Abbot, B. P., Abbott, R. and Abbott, T. D. et al. 2017. GW170814: A three-detector observation of gravitational waves from a binary black hole coalescence. *Physical Review Letters* 119: 141101.

Abend, S., Gebbe, M. and Gersemann, M. et al. 2016. Atom-chip fountain gravimeter. *Physical Review Letters* 117: 203003.

Acernese F., Athagos, M. and Agatsuma, K. et al. 2014. Advanced Virgo: a second-generation interferometric gravitational wave detector. *Classical and Quantum Gravity* 32: 024001.

Amaro-Seoane, P., Audley, H. and Babak, S. et al. 2017. Laser interferometer space antenna. arXiv:1702.00786.

Armano, H., Audley, H. and Baird, J. et al. 2018. Beyond the required LISA free-fall performance: new LISA pathfinder results down to 20 μHz. *Physical Review Letters* 120: 06011.

Aso, Y., Muchimura, Y. and Somiya, K. et al. 2013. Interferometer design of the KAGRA gravitational wave detector. *Physical Review D* 88: 043007.

Biedermann, G. W., Wu, G. W., Deslauriers, L., Roy, S., Mahadeswaraswamy, C. and Kasevich, M. A. 2015. Testing gravity with cold-atom interferometers. *Physical Review A* 91: 033629.

Cavendish, H. 1798. Experiments to determine the density of the earth. *Philosophical Transactions of the Royal Society of London* 88: 469–526.

D'Agostino, G., Desogus, S. and Germak, A. et al. 2008. The new IMGC-02 transportable absolute gravimeter: measurement apparatus and applications in geophysics and volcanology. *Annals of Geophysics* 51: 39–49.

Dooley, K. L., Leong, J. R. and Adams, T. et al. 2016. GEO 600 and the GEO-HF upgrade program: successes and challenges. *Classical and Quantum Gravity* 33: 075009.

Ferrari, G., Poli, N., Sorrentino, F. and Tino, G. M. 2006. Long-lived Bloch oscillations with bosonic Sr atoms and application to gravity measurement at the micrometer scale. *Physical Review Letters* 97: 060402.

Fixler, J. B., Foster, G. B., McGuirk, J. M. and Kasevich, M. A. 2007. Atom interferometer measurement of the Newtonian constant of gravity. *Science* 315, 74–77.

Gillot, P., Francis, O., Landragin, A., Pererira Dos Santos, F. and Merlet, S. 2014. Stability comparison of two absolute gravimeters: optical versus atomic interferometers. *Metrologia* 51: L15–L17.

Geiger, R., Amand, L. and Bertoldi, A. et al. 2015. Matter-wave laser Interferometric Gravitation Antenna (MIGA): New perspectives for fundamental physics and geosciences. arXiv:1505.07137.

Gundlach, J. H. and Merkowitz, S. M. 2000. Measurement of Newton's constant using a torsion balance with angular acceleration feedback. *Physical Review Letters* 85: 2869.

Hu, H., Svitlov, S., Rothleitner, C., Schafer, J., Zhang, J. and Wang, L. J. 2010. Improvements of the MPG-2 transportable absolute ballistic gravimeter. *Metrologia* 47: 575.

Hu, L., Poli, N., Salvi, L. and Tino, G. M. 2017. Atom interferometry with the Sr optical clock transition. *Physical Review Letters* 119, 263601.

Lamporesi, G., Bertoldi, A., Cacciapuoti, L., Prevedelli, M. and Tino, G. M. 2008. Determination of the Newtonian gravitational constant using atom interferometry. *Physical Review Letters* 100: 050801.

Martynov, D. V., Hall, E. D. and Abbott, B. P. et al. 2016. Sensitivity of the advanced LIGO detectors at the beginning of gravitational wave astronomy. *Physical Review D* 93: 112004.

Mazzoni, T., Zhang, X., Del Aguila, R., Salvi, L., Poli, N. and Tino, G. M. 2015. Large-momentum-transfer Bragg interferometer with strontium atoms. *Physical Review A* 92: 053619.

McGuirk, J. M., Foster, G. T., Fixler, J. B., Snadden, M. J. and Kasevich, M. A. 2002. Sensitive absolute-gravity gradiometry using atom interferometry. *Physical Review A* 65: 033608.

Mohr, P. J., Taylor, B. N. and Newell, D. B. 2012. CODATA recommended values of the fundamental physical constants: 2010. *Reviews of Modern Physics* 84: 1527.

Niebauer, T. M., Sasagawa, G. S., Faller, J. E., Hilt, R. and Klopping, F. 1995. A new generation of absolute gravimeters. *Metrologia* 32: 159–180.

Parks, H. V. and Faller, J. E. 2010. Simple pendulum determination of the gravitational constant. *Physical Review Letters* 105: 110801.

Peters, A, Chung, K. Y. and Chu, S. 1999. Measurement of gravitational acceleration by dropping atoms. *Nature* 400: 849–852.

Peters, A, Chung, K. Y. and Chu, S. 2001. High-precision gravity measurements using atom interferometry. *Metrologia* 38: 25.

Rosi, G., Sorrentino, F., Cacciapuoti, L., Prevedelli, M. and Tino, G. M. 2014. Precision measurement of the Newtonian gravitational constant using cold atoms. *Nature* 510: 518–521.

Rosi, G., Cacciapuoti, L., Sorrentino, F., Menchetti, M., Prevedelli, M. and Tino, G. M. 2015. Measurement of the gravity-field curvature by atom interferometry. *Physical Review Letters* 114: 013001.

Rothleitner, C. and Francis, O. 2014. Measuring the Newtonian constant of gravitation with a differential free-fall gradiometer: a feasibility study. *Review of Scientific Instruments* 85: 044501.

Schwarz, J. P., Robertson, D. S., Niebauer, T. M. and Faller, J. E. 1998. A free-fall determination of the Newtonian constant of gravity. *Science* 282: 2230–2234.

Vitushkin, A. L. and Faller, J. E. 2002. Development and investigation of a compact transportable absolute ballistic gravimeter. *Measurement Techniques* 45: 893–900.

12

Satellite Laser Ranging

José C. Rodríguez and Graham M. Appleby

CONTENTS

12.1 Introduction 181
12.2 Working Principle 182
  12.2.1 General Description 182
  12.2.2 Components 183
    12.2.2.1 Ground Segment 183
    12.2.2.2 Space Segment 186
12.3 Operations 188
  12.3.1 Signal Strength 188
    12.3.1.1 Track Detection 188
    12.3.1.2 Data Screening 189
  12.3.2 Ranging Policy 189
  12.3.3 Calibration 191
  12.3.4 Safety 191
  12.3.5 Global Network 191
12.4 Analysis 192
  12.4.1 Centre of Mass Corrections 193
  12.4.2 Accuracy and Quality Control 193
12.5 Applications 193
  12.5.1 Terrestrial Reference Frame 193
  12.5.2 Support of Scientific Missions 194
  12.5.3 Tracking of GNSS Constellations 195
  12.5.4 Space Debris 195
  12.5.5 Time Transfer 195
  12.5.6 Lunar Laser Ranging 196
12.6 Concluding Remarks and Future Trends 196
References 196
Further Reading 198

12.1 Introduction

Satellite laser ranging (SLR) began as a concept in 1962 when Plotkin of the Goddard Space Flight Center in Maryland, USA, first proposed the development of accurate laser ranging to retroreflectors on orbiting spacecraft in order to improve geodetic information (Plotkin 1964). At that time optical and radar tracking of satellites was being used to yield tracking station coordinates at a level of accuracy of only about 100 m. The interesting geophysical processes that deform the solid Earth, such as Earth tides and plate tectonic motions, were understood at that time to be affecting station coordinates at the level of only a few centimetres over timescales varying from sub-daily to several years. It was clear that in order to challenge the theoretical framework the measurement techniques would have to reach a level of accuracy of a few centimetres and the observations would need to be carried out over many years.

The first SLR observations, of the Beacon Explorer-B spacecraft in 1964, achieved metre-level precision in range and showed that the technique was viable and potentially capable of reaching the required centimetric precision. These experiments prompted NASA to place retroreflectors on the GEOS I and II spacecraft and on the Moon, and to begin the development of more SLR stations. In 1975 France launched the first geodetic satellite STARLETTE into a relatively low 950 km orbit, followed in 1976 when NASA launched its Laser Geodynamic Satellite LAGEOS into a near circular, near polar orbit at a height of 6000 km. Both these satellites are inert, dense spheres, uniformly encrusted with retro-reflecting cube-corners specifically designed to reflect laser light back to the emitting tracking station.

In 1979 NASA created the Crustal Dynamics Project with the aim of 'developing laser ranging and very long baseline interferometry (VLBI) systems to obtain relative positions with ±2 cm accuracy, to define directions with respect to the inertial reference with a 0.001 arcsecond accuracy to monitor relative rates of motion of different parts of the Earth's crust well enough to infer irregularities in plate tectonic motions, and to monitor the wobbles and rotational variations to infer their excitations and dampings, as well as to determine accurately the orbits of the distant satellites (higher than 6000 km altitude)' (Bosworth et al. 1993). This project formalized and consolidated the cooperative efforts between scientists and engineers, which were already underway in several countries, and other independent groups worldwide began to design and build their own SLR systems.

Nowadays SLR is one of four space geodetic techniques that combined are employed to realize a global terrestrial reference frame that provides the positional infrastructure required to link in both space and time scientific observations taken on, and above, the Earth's surface. The current SLR network comprises over 40 tracking stations globally distributed run by scientific institutions from different countries. Together, under the auspice of the International Laser Ranging Service (ILRS), the SLR network provides ranging measurements of unsurpassed accuracy to various kinds of artificial satellites, from low Earth orbiting spacecraft with scientific payloads to communication satellites in geosynchronous orbits. In addition, lunar capable stations routinely perform laser ranging measurements to retroreflectors located on the surface of the Moon.

SLR data provide the coordinates of the Earth's centre of mass; defines the scale of the terrestrial reference frame together with VLBI; provides the first term of the Earth's gravity field series harmonic expansion; contributes to the realization of the global reference frame with station positional information and Earth rotation data; it is employed as a tool to test general relativity; it is the benchmark against which orbits for Earth observation, experimental, and communication satellites are compared to and their models refined and validated against.

12.2 Working Principle

12.2.1 General Description

The SLR technique is, in principle, very simple. Short pulses of laser light are emitted from a telescope that is driven to follow the predicted position of an orbiting satellite, with the time of emission accurately recorded by a timing device accurate and precise to several picoseconds. Light pulses reach the cube-corner retroreflectors mounted onboard the satellite and are directed back towards the telescope. Upon receipt of the reflected laser light, a high-speed detector generates an electronic signal that triggers the timer, thus recording the round-trip time-of-flight. Half this time-of-flight multiplied by the speed of light, plus some small corrections, gives the range to the satellite at the emission time of the laser pulse, known as the epoch of the observation. Repeated measurements performed while the satellite is visible from the tracking station provide very accurate orbital information over that pass segment. The principal components of a tracking station are shown schematically in Figure 12.1.

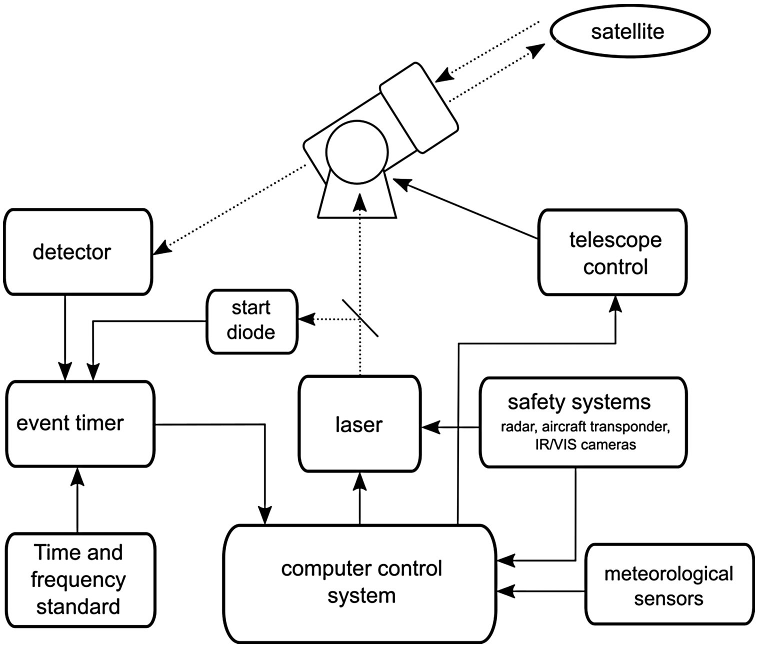


FIGURE 12.1 A computer-controlled telescope tracks a satellite carrying retroreflectors. The time-of-flight of laser pulses to and from the satellite is measured to high precision to determine the satellite range and the measurement is time-stamped with an accurate epoch.

12.2.2 Components

12.2.2.1 Ground Segment

12.2.2.1.1 Lasers

Given the principle of SLR, namely to measure as precisely and accurately as possible round-trip transit times from an observatory to an orbiting satellite, the characteristics of the laser are of vital importance in this endeavour. The requirement is a pulsed laser of sufficient power to reach the satellites of interest, whose pulse length must be as short as possible to minimize uncertainty in the measurement. The evolution in precision of SLR measurements has therefore closely followed the technological development of short-pulsed lasers, covered in detail elsewhere in this volume. The early experiments (1960–1970) with satellite and lunar ranging used, of necessity, mainly ruby lasers. Q-switching was carried out by a rapidly rotating cavity mirror and the system delivered typically 2 J of energy in single pulses a few nanoseconds in duration. The pulse duration was determined by practical limitations on the rotation rate of the mirror, and repetition rates remained very low (about 0.2 Hz) in order to reduce excessive heating in the laser rod. The measured round-trip time could not be resolved more precisely than the pulse length, leading to inferred satellite ranges of precision about 50 cm.

With the development of non-mechanical active switching techniques such as electro- and acousto-optic Q-switching and passive Q-switching using saturable absorbers came the ability to generate sub-nanosecond laser pulses, leading to measurement precisions of 10–20 cm. A quantum leap in precision came as a result of the development of reliable mode-locking systems. In a mode-locked system, a train of extremely narrow, equally spaced pulses is produced during each laser cycle, instead of the single Q-switched pulse. To obtain high-power, short-duration pulses, the system should be designed to have a large number of longitudinal modes, which implies a system with a broad laser transition and a long laser cavity. When a saturable dye is used to mode lock the laser, then the laser is simultaneously Q-switched, resulting in a train of pulses, each of duration 10–50 ps, in an envelope that may be several hundred nanoseconds long. Each pulse will have a peak power measured in GW. A laser system particularly suited to operating in this way is the Nd:YAG, the most popular choice in the previous generation of SLR systems, with repetition rates in the region of 10 Hz. Its bandwidth of some 120 GHz sets a lower limit on pulse width of about 10 ps, but in practice, to avoid excessive damage of optical components; this was normally increased to at least 30 ps using an etalon in the cavity.

Further developments in solid-state laser technology in the 1990s would enable the onset of a new generation of SLR systems in the next decade. Diode-pumped solid-state (DPSS) lasers have replaced flashlamp systems in many applications, thanks to the several advantages they present (higher efficiency and hence lower operating cost, lower maintenance, higher durability, compactness). In the field of ultrafast, high-power applications, DPSS lasers are now ubiquitous, based on lasing materials such as Nd:YAG and Nd:YVO4. The new generation of SLR systems employs high repetition rate (>1 kHz), ultra-short pulsed lasers (10 ps). These lasers present evident advantages such as enabling data volumes and acquisition rates up to two orders of magnitude superior to the legacy systems, as well as higher measurement precision due to their short pulse length and greater number of successful observations available. In addition, they have made possible new data reduction strategies, as the high repetition rates allow the discrimination of single cube-corners within the satellite retroreflector array. The intrinsic precision of the new systems means that the laser sub-system is no longer a limiting factor in the SLR technique. As a downside, these high-repetition lasers present a lower pulse energy, typically a few hundred µJ at most, severely limiting the single-shot probability of detection on all but the most easily accessible satellite targets. Therefore, it becomes more critical to ensure that the optical, detection and tracking systems operate optimally to maximize detection probability and minimize energy losses.

The output wavelength of the most popular lasing materials employed in SLR laser systems, 1.06 µm, is far from optimum for laser ranging, since high-speed detectors with high quantum efficiency are optimized for the visible. It is a common practice, therefore, to double the frequency of the light in the laser pulses so that they are emitted at a final wavelength of 0.53 µm in the green. Frequency doubling, or second harmonic generation, can be achieved with efficiencies approaching 50% using crystals of KDP (potassium dihydrogen phosphate) or lithium niobate. However, there is a small number of stations that operate, or have done so in the past, at frequencies other than green, and recent developments in detector technology have greatly alleviated previous constraints in wavelength. This opens up new and interesting possibilities in system configuration, as direct use of the infrared output avoids the 50% losses incurred in frequency doubling, and the atmosphere has a higher transparency in this region than in the visible. For instance, laser ranging using infrared wavelengths has proved very successful for lunar ranging at the French station in Grasse (Courde et al. 2017).

12.2.2.1.2 Telescopes

In order to range to Earth-orbiting reflectors, whether mounted on artificial satellites or placed on the surface of the Moon, a computer-controlled telescope is required to direct the laser pulses accurately towards the targets and receive the reflected photons on their return. Telescopes for lunar laser ranging (LLR) differ little from standard astronomical telescopes in their optical and tracking characteristics: the required tracking rate is relatively slow and high-quality optics deliver the outgoing light with minimum divergence, ensuring the maximum photon density at the reflector in space. Indeed, the earliest ranging experiments to the Moon were made with the 2.7 m reflector at the McDonald Observatory, Texas, USA.

On the other hand, artificial satellites orbit the Earth much faster than the Moon does. For instance, communications satellites (e.g. GLONASS, Galileo, and GPS) orbiting 20 000 km above the Earth's surface take about 12 h to complete one revolution and are therefore visible from a given ground station for several hours in the course of each orbit. At the other end of the scale, satellites only 400 km above the Earth (approximately the same height as the International Space Station) complete a revolution in little over 90 min: they are only visible to stations during 2 or 3 min of each orbit. Telescopes capable of following these satellites must track at speeds of up to 2° s^{-1} while maintaining an absolute pointing accuracy of a few arcseconds. Because of the altitude-azimuth configuration of most modern telescopes (two independent rotational axes, a vertical one for azimuthal motion, and a horizontal one perpendicular to it for elevation), the azimuth velocity required to track satellites that pass near the local zenith can be very large. Consequently, SLR telescopes are designed to slew at speeds of typically 20° s^{-1}, with high acceleration capabilities. Nevertheless, for fast-moving low Earth orbiters (LEOs) there often is a blind area around the zenith, a few degrees in diameter, where the telescope drives cannot keep up and satellite tracking is momentarily lost.

For all types of telescopes the ability to point to a position anywhere in the sky to within a few seconds of arc is achieved by using a mathematical model to transform the coordinates read from the telescope encoders to the sky coordinate system used to describe the satellite motion. This is performed periodically by observing a set of stars (which have very accurately known catalogue positions) distributed over the entire visible hemisphere. A six-to-ten parameter model may then be fitted to the differences between the observed and catalogue coordinates, which is used subsequently in real time to correct the raw encoder coordinates to sky coordinates. These differences arise from small imperfections in the mechanical construction of the telescope and its setting up: the altitude and azimuth axes will not be precisely at right angles, nor will they be exactly horizontal or aligned with the geographical north; the telescope tube may flex differentially under gravity as it moves in altitude; and the encoders on the telescope axes will have zero-point offsets. Adjustments are also applied to the elevation in real time to correct for refraction in the Earth's atmosphere, which is a function of zenith distance, barometric pressure, temperature, and relative humidity (meteorological data are readily available since they are also mandatory to perform range corrections).

Laser ranging telescopes also differ from their astronomical counterparts in that they are regularly used in broad daylight. Because of the high tracking rates required not all domes are capable of rotating at sufficient speed to follow the telescope, so many SLR instruments are housed in non-rotating domes where most of the roof opens to the sky. During daylight hours this means that parts of the structure of the mount may be exposed to full sunlight, causing differential expansion and potentially compromising the telescope pointing accuracy. Insulating the vulnerable parts or deriving separate pointing models for day and night afford a partial remedy, although the ideal solution is to operate with domes designed to minimize direct sunlight exposure. These have also the advantage of providing greater protection to the telescope and mount from strong wind conditions, which can be common in certain locations at high elevations or latitudes.

In terms of the arrangement of the receive and transmit optical axes, there are two main categories of SLR telescopes: those employing the same optical path for both sending and detecting the laser light, usually termed transmit/receive or monostatic systems, and those with separate, co-mounted telescopes (bi-static systems). A much less common set-up consists in having two entirely separate, individually mounted and controlled telescopes, as in Potsdam station (Germany). A concept expanding on this has been proposed recently, consisting in the deployment in different locations of receive-only systems, which would detect the laser echoes from a single, remote transmit/receive site. This would enable the collection of more data at low cost and complexity, which could see some use in the emerging field of space debris tracking. Tests demonstrating this technique have already been performed, whereby laser pulses from a station in Graz (Austria) to a defunct satellite (Envisat, ESA) were successfully detected by distant stations in Zimmerwald (Switzerland), Wettzell (Germany), and Herstmonceux (UK).

The need for physical stability and a regulated environment to achieve a consistent performance from the laser and timing electronics has traditionally dictated that the laser be mounted on a solid optical bench in a temperature-controlled room or enclosure adjacent to or below the telescope structure. The light pulses are brought to the telescope *via* a series of flat mirrors forming a coudé chain, all coated for maximum reflectivity at the employed laser wavelength. In monostatic systems the final mirror directs the light on to the optical axis towards the secondary mirror of the telescope. Reflection from the secondary fills the main mirror that then sends the light pulses out into space. The orientations of the mirrors of the coudé chain must be carefully adjusted so that the laser beam lies along the optical axis regardless of telescope orientation. Detectors are often mounted at the Cassegrain focus of the telescope, although part of the coudé chain may also be used to bring the return pulse to a fixed detector. Whichever method is adopted, at some point an optical component has to be moved so that the incoming light does not return to the laser. A common scheme is to employ a pierced rotating flat as the final element of the coudé chain; the timing of the rotation has to be carefully controlled so that a reflecting section is in position on the optical axis for the outgoing pulse, and the hole is in position for the return. This requirement can present severe problems for very low satellites, and for ranging to ground targets, where the short time-of-flight does not give enough time to remove the coudé element before the arrival of the return signal.

Where dual telescopes are present on the same mount, the coudé chain feeds only the emitter telescope, usually a small refractor, leaving the main telescope free to fulfil its receiving role without interruption. In this arrangement it is necessary to maintain the parallelism of the optical axes of the two telescopes, and again to adjust the coudé mirrors to deliver the outgoing light centrally in the emitter telescope and along its optical axis. The placement of a pair of lenses, one concave and one convex, in the emitter telescope allows variation in the divergence of the outgoing beam by simply moving one of the lenses longitudinally along the optical axis. While this is not a necessity, expanding the divergence may be used to relax the pointing precision required for tracking in the case of satellites with strong return signals, to shorten the acquisition time, and to control the delivered energy on target (e.g. tracking of Sentinel-3, an Earth observation mission by ESA, is only allowed to those stations that can ensure a limited energy flux to avoid potential damage to sensitive on-board detectors).

An iris assembly in the focal plane allows the field-of-view to be limited so as to reduce incident sky noise during daytime observing. Spectral filtering may further reduce sky noise. In most systems light from the telescope is first collimated and then passed through a dichroic mirror chosen to transmit only a narrow band centred on the wavelength of the laser light. Such a dichroic may be configured so that all the non-laser light is reflected at right angles to the telescope axis where, at night, it becomes available to other detection equipment (e.g. low-light cameras or photometers). The light transmitted through the dichroic is directed to the detection package, normally after passing through some additional filtering components to reduce the band pass to typically a few Angstroms and hence minimize the presence of noise. These narrow filters are very effective at suppressing noise events, but at considerable transmission losses (50% typical) of the returned laser light, which can be the determining factor for success or failure in daytime observing of distant satellites.

12.2.2.1.3 Detectors

The function of the detector in laser ranging systems is to respond to an incoming light pulse and produce a sharply defined electrical signal that will accurately and consistently trigger the timer to record the instant of arrival. By contrast with the situation at the start diode, where each laser pulse may contain upwards of 10^{17} photons, the return signal is hugely diminished and may indeed consist of only a single photon. Photomultiplier tubes (PMTs), employed in the first generations of SLR systems, have been mostly superseded by newer devices that offer many advantages over them. PMTs are used in many low-light level applications in astronomy, but almost always in the context of photon counting for intensity measurements rather than for precise timing. A weak light signal arriving at the photocathode initiates an electron cascade in the device and the electrical pulse that emerges has a very fast rise-time. It is the constancy of the time interval between the arrival of light at the photocathode and some measurable point on the leading edge of the electrical pulse from the detector that ultimately determines the suitability for timing. In PMTs the number of photons simultaneously hitting the photocathode, or the particular path taken by the electron cascade, can both affect this interval and thereby add considerable jitter to the overall time-of-flight measurement. To alleviate to a large degree the dependency of the timing jitter on the amplitude of the return pulse, it is a common practice to pass the output signal from the PMT through a constant fraction discriminator. In this way, the timing jitter may be reduced to a few hundred picoseconds, equivalent to a range uncertainty of a few centimetres.

Modern ranging systems have largely abandoned PMTs in favour of micro-channel plate (MCP) detectors (in combination with constant fraction discriminators) or avalanche photodiode detectors working in Geiger mode. In MCP detectors, accelerated photoelectrons generated at the photocathode enter a plate containing a dense array of cylindrical channels a few microns in size. As photoelectrons bounce off the walls of the channels, more electrons are generated providing final gains in excess of 10^6. The short length of the microchannels makes the path followed by the generated electrons much shorter and better defined, resulting in very small time jitter compared to PMTs (Degnan 1985). Single-photon avalanche diodes (SPADs) are particularly well suited to low-energy ranging systems (Lightstone and McIntyre 1981). These solid-state devices are both more compact and more robust than their PMT counterparts but, with a photosensitive area an order of magnitude smaller, optical alignment is much more critical. Although the amplification mechanisms are similar, both the transit time for the electron cascade and its time spread are reduced in the smaller structures, so that the outgoing electrical pulse has a faster rise-time and reduced jitter. Cooling the detector chip in avalanche diode detectors can significantly improve the component of jitter arising from intrinsic electronic noise. However, the processing electronics of detectors mounted on the telescope are often exposed to the diurnal cycle of ambient temperature change and can result in apparent range dependence at the level of a few millimetres. Frequent calibration measurements to a local terrestrial target solve this problem.

A critical endeavour in the detection process is the elimination of time-walk, which is the systematic dependence of the time marker on the energy of the input pulse. A higher energy input signal produces a higher amplitude output pulse with a faster rise-time; such a pulse reaches the discriminator level earlier and thus records a slightly different time. The time difference between pulses of 1 and 100 photons hitting the detector may be equivalent to 10 mm or more in the measured range. One approach to minimizing this effect is to restrict the return energy to a known regime, usually single photons, thereby ensuring that the time-walk is identical for all measurements. Another solution is to determine the time-walk characteristic for a given device, provide extra circuitry to detect the rise-time slope of the outgoing pulse and add a suitable, slope-related time delay so that all outgoing pulses are delivered as if the input was a single photon. Understanding and compensating such effects have been major factors in improving precision and stability of the SLR technique.

12.2.2.1.4 Timers

The ultimate success for the technique of laser ranging lies in the accuracy with which it is possible to measure the time-of-flight of a laser pulse to the target satellite and back. In geodetic applications it is also necessary to record the exact epoch at which the measurement is made: then the relative positions of the moving satellite and the observing station on the rotating Earth can be established. Given the objective of making range measurements to mm-level precision, it is straight-forward to estimate the precision with which the epoch of the observations must be time tagged. The velocity of a satellite in circular orbit around the Earth at an altitude of 5860 km (the altitude of the geodetic satellite LAGEOS) is about 5.7 km s^{-1} or 5.7×10^6 mm s^{-1}. Thus, observations should be recorded to a precision of better than 175 ns. For lower satellites, the requirements would be more stringent. Historically, epochs at stations were derived either from a local high-quality frequency standard or directly from a local time service, if available. Nowadays GPS receivers have largely replaced the local frequency standards, making timing accuracies to better than 100 ns reliably available at any station anywhere in the world. The timing outputs from a GPS receiver consist of a 1 pps (pulse per second) for synchronization to UT and a 10 MHz sine wave, which can be used to drive the system clock. The range measurement itself must, of course, be made to the required precision of a few millimetres, which, in terms of two-way measured time-of-flight, is equivalent to a precision of better than about 10 ps. In order to convert this measurement precision into measurement accuracy, range calibration measurements are frequently performed to a terrestrial target, whose accurate distance from a reference point in the ranging telescope has been independently surveyed. The calibration measurements are used to determine the sum of the internal electronic and optical-path delays in the ranging system, which is then removed from each raw range observation during a pre-processing stage of data reduction.

Two different kinds of timing devices are employed to time tag the epochs of observations and measure the time-of-flight intervals of the laser pulses. The oldest, still employed in many stations, uses a clock with a 100 ns resolution and a picosecond interval timer to perform these two tasks separately. Time interval units (TIU) achieved accuracies of 100 ps with 20 ps time resolution, and the maximum range they could measure was limited so that for high-altitude targets the number of concurrent in-flight laser pulses had to be reduced. TIUs may suffer from non-linearity effects and instabilities in the internal oscillator, which translates into both random and systematic range-dependent errors up to the centimetre level. Careful calibration of these devices is required for optimal operation, which ideally involves stacking more than one unit to extract the average timings of different instruments. New and upgraded stations now employ event timer devices instead of time interval counters. Event timers use a picosecond resolution clock to record the precise epochs of both the firing and detection events, which are then subtracted to obtain the time-of-flight. Event timers have vastly superior precision, linearity, and time stability characteristics than interval counters. In these devices the time range is practically unlimited, and the measurement rate is greater than 10 kHz. For any station employing a modern laser with firing rates superior to 10 Hz, the use of an event timer is mandatory.

12.2.2.2 Space Segment

12.2.2.2.1 Retroreflectors

Regular SLR operations require that satellites be equipped with one, or more, laser retroreflectors (in the latter case we speak of laser retroreflector arrays). In terrestrial applications, such as surveying, retroreflectors are perfect cube-corners: a glass block with three of its surfaces mutually perpendicular. A light ray entering *via* the front surface is reflected three times adding to a total angle of 180° and returns on a path parallel to the incoming direction. The reflector performs this function over a wide range of acceptance angles (usually a cone of half-angle about 30° to 40° whose axis is perpendicular to the front face of the reflector) and the length of the light path through the glass is the same for all rays. Such reflectors, therefore, provide a very effective means of getting strong return signals at the point where the light beam originates, giving a consistent range measurement independent of where or at what angle the light enters. In comparison with, e.g., a polished sphere at the same distance, retroreflectors act as a powerful amplifier.

Obtaining returns from retroreflectors in space poses two additional problems compared to ground-only applications. First, both the target and the observer are in motion; second, the vast distances involved result in huge power losses. The relative motion of telescope and target means that, at the moment the laser is fired, the telescope must point to the direction of the satellite at the time the laser pulse arrives (*point ahead*); this adjustment may be computed in real time from the satellite orbit predictions. The fact that the observing station is fixed on the surface of a rotating Earth poses additional complications. Laser light reflected at the satellite will travel back down the same path to a point on the Earth's surface displaced from the receiver telescope. In practice, this means that the telescope is not at the centre of the reflected light pulse, which is the point where the energy flux is maximum. As retroreflectors in space act as small apertures, the reflected light generates a diffraction pattern which in the far field, for a single, perfect retroreflector with a circular entrance window and in a direction perpendicular to it is the Airy pattern. In most situations, the diffraction patterns of reflected pulses are more complex, for usually several retroreflectors are involved and the laser beam is not perpendicular to their front faces. However, in broad terms, the intensity is smaller as the distance to the centre of the diffraction pattern increases. In order to maximize the received signal retroreflectors are manufactured so that the dihedral angles of their internal faces are not perfectly perpendicular. This alters the diffraction pattern so that less energy is concentrated in its centre region, while the intensity off-axis is increased. Typical dihedral angle offsets employed are in the order of 0.5–4 s of arc.

A relatively recent innovation in retroreflector technology is based on the concept of the Luneburg lens. Light incident on a perfect, solid sphere of transparent glass of refractive index (RI) precisely 2.0 is refracted inside the sphere, reflected from the inside surface opposite the point of entry, refracted again, and finally emerges parallel to the incident beam. In order to obtain an effective RI=2.0 in a large, homogeneous mass, different materials are nested concentrically to achieve the same effect. The first realization of this concept, a 6 cm diameter sphere attached to a METEOR-3 satellite launched in December 2001, was extensively tracked by the global SLR network. In 2009 BLITS (Ball Lens in The Space), a larger Luneburg sphere about 8.5 cm diameter, was launched by the Russian space agency ROSCOSMOS. In this case the satellite is the retroreflector, which fulfils the goal of offering a target exempt of satellite signature effects.

12.2.2.2.2 Satellite Missions

Satellites routinely tracked by the worldwide network of SLR stations fall into three main classes: geodetic, applications, and navigational.

2.2.2.2.1 Geodetic spheres Geodetic satellites are small, inert spheres containing a high-density core and a surface uniformly covered with cube-corner retroreflectors. This design minimizes the area-to-mass ratio and hence accelerations due to non-gravitational forces such as atmospheric drag, solar radiation pressure, and albedo. Geodetic satellites that are regularly tracked by the global network are listed in Table 12.1. All of these are in near-circular orbits and have diameters ranging from 20 cm to over 2 m. Figure 12.2 shows the latest geodetic sphere launched, LARES, a mission of the Italian Space Agency put in orbit in 2012. The satellite has a diameter of 0.182 m and carries 92 retroreflectors distributed over its surface, which are represented by the dots in the figure. 2140 of the reflectors are optimized to reflect the conventional green laser light used by SLR systems, and six of them are germanium reflectors optimized for infra-red tracking. The principal applications addressed by observations of geodetic satellites depend upon their orbital heights. The LAGEOS satellites are extensively used for global terrestrial reference frame determination in order to study crustal motions, Earth rotation and for determination of, and monitoring changes in, the lowest frequency terms in the models used to represent the Earth's gravity field. Lower satellites such as Starlette, Stella, Ajiasi and LARES are better suited to the determination of higher frequency terms in the expansion of the Earth's gravitational potential, studying ocean tides and modelling variations in atmospheric density.

2.2.2.2.2 Applications Satellites The applications satellites are large, irregularly shaped objects whose payload typically include various remote-sensing devices, batteries and solar arrays for power generation, attitude control systems, communications equipment, and certain manoeuvrability capabilities. Examples of scientific missions that are being tracked at present by the SLR network include the joint French-USA altimetry missions Jason-2 and Jason-3, NASA's GRACE-A and GRACE-B for gravity field determination, ESA's Swarm constellation for geomagnetism studies, and the Chinese altimeter satellite HY-2A. Depending on the specific mission, SLR observations are used for validation of the orbits generated from their primary navigation systems (GNSS and/or DORIS) or contribute directly to their precise orbit determination (POD). Laser ranges, especially at high elevations, are treated as the ground truth against which different POD modelling and strategies are compared. In the past, laser ranging has provided the main tracking system for POD after failure of other systems; the GFO mission was effectively rescued by including a laser retroreflector array onboard: when shortly after launch it was discovered that the GPS receivers could not track more than one satellite simultaneously, it was possible to compute precise orbits relying on SLR, Doppler tracking, and altimeter cross-over data.

2.2.2.2.3 Navigation Constellations Global navigation satellite systems (GNSS) are satellite constellations that provide signals at a global coverage employed by ground and space receivers to obtain timing and positional information. The role of SLR in supporting GNSS operations is to provide an independent check on the quality of the orbit determination derived by the radiometric tracking systems. Beyond a pure quality control check, this information is also used to assess, test, and develop the dynamic modelling of the satellite orbits, e.g. solar radiation pressure, which amounts to the biggest non-conservative force acting on these satellites. All the

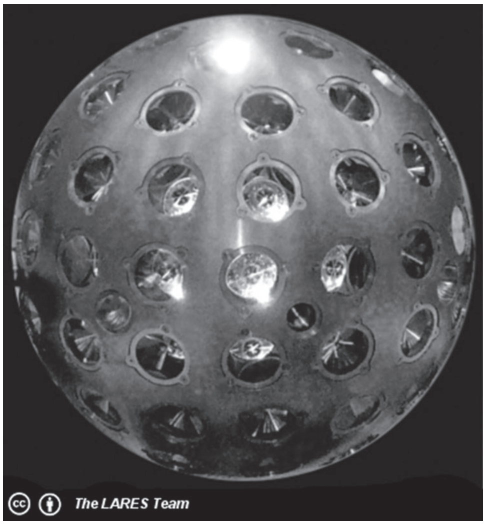


FIGURE 12.2 Satellite LARES, a tungsten sphere covered with 92 cube corner retroreflectors, primarily designed to perform General Relativity tests. LARES orbits the Earth at 1440 km altitude, completing a revolution in approximately 114 min. LARES is the most dense object in orbit in the solar system. Picture by the LARES Team (Distributed under license Creative Commons CC BY 3.0. https://creativecommons.org/licenses/by/3.0/).

TABLE 12.1

Spherical Geodetic Satellites Currently Tracked by the ILRS Network

	Height (km)	Diam (cm)	Mass (kg)	Inclination (°)	Period (min)	Number of Retros	Launch Date
Larets	690	24	23.3	97.8	98	60	2003
Starlette	812	24	47	49.8	104	60	1975
Stella	804	24	48	98.6	101	60	1993
LARES	1440	36.4	396.8	69.6	114	92	2012
Ajisai	1500	215	685	50.0	116	1436	1986
LAGEOS-1	5860	60	407	109.9	225	426	1976
LAGEOS-2	5620	60	405	52.7	223	426	1992
Etalon-1	19120	129	1415	64.9	676	2146	1989
Etalon-2	19210	129	1415	65.5	675	2146	1989

spacecrafts of the Russian GLONASS, the European Galileo constellation, Chinese BeiDou, and Indian NAVIC, carry laser retroreflector arrays. While so far only two GPS satellites have been equipped with retroreflectors, the next generation of spacecraft will include them.

Because of the different orbital characteristics of the numerous satellites tracked by laser ranging, the times of visibility during which observations from a given tracking station can be attempted vary from a few minutes for LEOs to several hours for the high-altitude GNSS spacecraft. Satellite visibility may even be permanent or nearly so, as it is the case with the geosynchronous orbits employed in some GNSS systems (BeiDou, NAVIC, QZS). Tracking station schedules must therefore prioritize their activities to ensure good coverage of primary satellites such as LAGEOS and other scientific missions. As modern systems have a short acquisition time and are capable of switching tracking in a quick fashion, the capability of the network has not been exhausted even though the current list of satellites routinely tracked approaches 100 objects.

12.3 Operations

12.3.1 Signal Strength

Laser pulses emitted from a ground station will typically have Gaussian spatial and temporal characteristics of a few centimetres (up to a few tens of cm) across and tens of picoseconds long (10–150 ps FWHM). Because of beam divergence, the laser pulses arrive at the satellite spread out to up to hundreds of metres in diameter. The photon energy incident at a reflector at range R from the emitter is proportional to R^{-2}. Similar considerations apply to the return energy at the telescope so that the overall return signal dependence with distance is a function of R^{-4}. The signal intensity depends mainly on the combination of energy in the emitted pulse, the divergence of the emitted beam, the size of the retroreflector array, and the collecting area of the receiving telescope. In more detail, the success or otherwise of SLR measurements is governed by the standard radar link equation (Degnan 1993), which provides an estimation of the expected average number of detected photoelectrons N_{pe} for each transmitted laser pulse:

$$N_{\mathrm{pe}} = \eta_{\mathrm{q}} E_{\mathrm{T}} \frac{\lambda}{hc} \eta_{\mathrm{T}} G_{\mathrm{T}} \sigma_{\mathrm{Sat}} \left(\frac{1}{4\pi R^2} \right)^2 A_{\mathrm{T}\eta\mathrm{R}} T_{\mathrm{A}}^2 T_{\mathrm{C}}^2 .$$

Here, η_{q} is the detector quantum efficiency, E_{T} is the energy of the laser pulse, η_{T} and η_{R} are the transmit and receive path efficiency, and A_{T} is the area of the receiver aperture. G_{T} is the gain of the transmitter, which may be determined as a function of pointing accuracy and beam divergence. T_{A} is the one-way transmission of the atmosphere and T_{C} is the one-way transmission of cirrus clouds if present. The satellite, at slant range R, has optical cross-section σ_{Sat}. For reference, the value of σ_{Sat} for a typical spherical geodetic satellite is of the order of 10^7 m^2 and for the small arrays on modern altimetry satellites an order of magnitude less. This equation has more value as an engineering tool to estimate the order of magnitude impact of different variables than as an analytical predictor of actual system performance. Some of the quantities required are difficult to estimate accurately (retroreflector cross-section, optical efficiencies, actual pointing errors) while others are time-varying (atmospheric conditions affecting laser propagation). It is informative to compute the detection performance for a given station under different scenarios; that is, keeping fixed those variables defining a specific system (detector quantum efficiency, laser energy, receiver telescope area, emitter gain) as well as those upon which ground stations have no control, i.e. satellite cross-section and distance. The expected return signals at elevation angles from 15° to 90° are shown in Figure 12.3 for both optimal conditions (high atmospheric transparency, absence of cirrus clouds, night time operations) and a severe observing scenario (very poor visibility, presence of cirrus clouds, day time ranging). The curves are normalized to the detection signal computed for good conditions at 90° of elevation. At high elevations (>60°), both curves behave similarly, with a performance under bad conditions about 14% of that expected under excellent ones. As satellite elevation decreases, and so the greater the air mass the laser pulses must traverse, the differences become dramatically pronounced. At 20° over the horizon and severe observing conditions, only 0.35% of the signal would be expected relative to the most optimal scenario. The change in signal intensity with satellite elevation under very poor conditions is about 4 orders of magnitude, a substantial dynamic range that tests the detection equipment and tracking stations ability to produce consistent, accurate time-of-flight measurements. To this end, many stations control, or at least limit, the intensity of the received signals in order to ensure a stable behaviour of the detection electronics, as well as effective, consistent system delay calibrations. An observing policy that takes this principle to its ultimate logical consequence is the so-called single-photon ranging. This consists in monitoring and controlling the detection rate of the laser pulses reflected off the satellite to sufficiently low levels so that, statistically, no more than a single photoelectron per laser shot on average is generated in the detector package. This strategy ensures the absence of return rate-dependent effects in the measured distances.

12.3.1.1 Track Detection

In daytime operations the recorded return signal from satellites may be contaminated by photons from the sky background. The noise level is minimized in three ways: by employing very narrow band filters in the light path (spectral filtering); by reducing the area of sky in the field of view with a small iris in the focal plane (spatial filtering); and by gating on the detector as close as possible to the expected time of arrival of satellite returns (temporal filtering). In addition, for systems working at a multi-photon return level, the detection electronics can be set to reject events below a certain energy threshold, which are predominantly single-photon noise detections. If the discriminator electronics are set to trigger only at, e.g. energies greater than three simultaneous photons, then nearly all noise events can be rejected. Of course, this requires that the satellite link is maintained at higher levels than the chosen threshold. For single-photon systems the satellite returns have to be extracted from the noise after detection. Each time the detector has

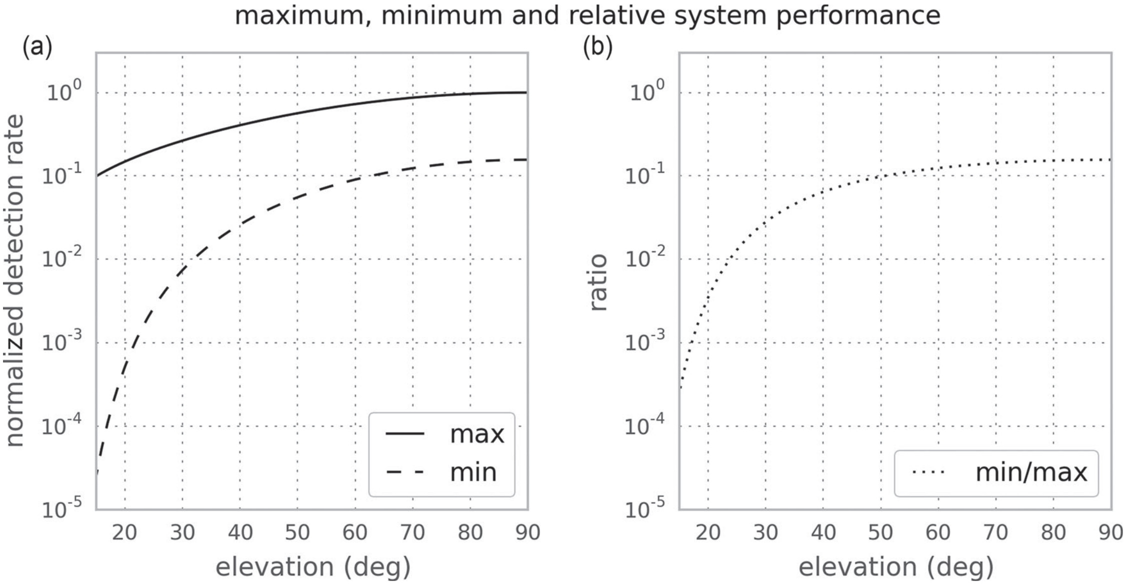


FIGURE 12.3 Relative intensities of returned laser pulses at different elevations under optimistic and pessimistic scenarios of observing conditions. The solid curve (a) shows a factor of 10 reduction in energy with elevation when atmospheric conditions are optimal. With poor conditions (dashes) the minimum reduction obtains at the zenith (~90% less), decreasing by 4 orders of magnitude at low elevations. The ratio between the normalized return energies at both scenarios is plotted on the (b).

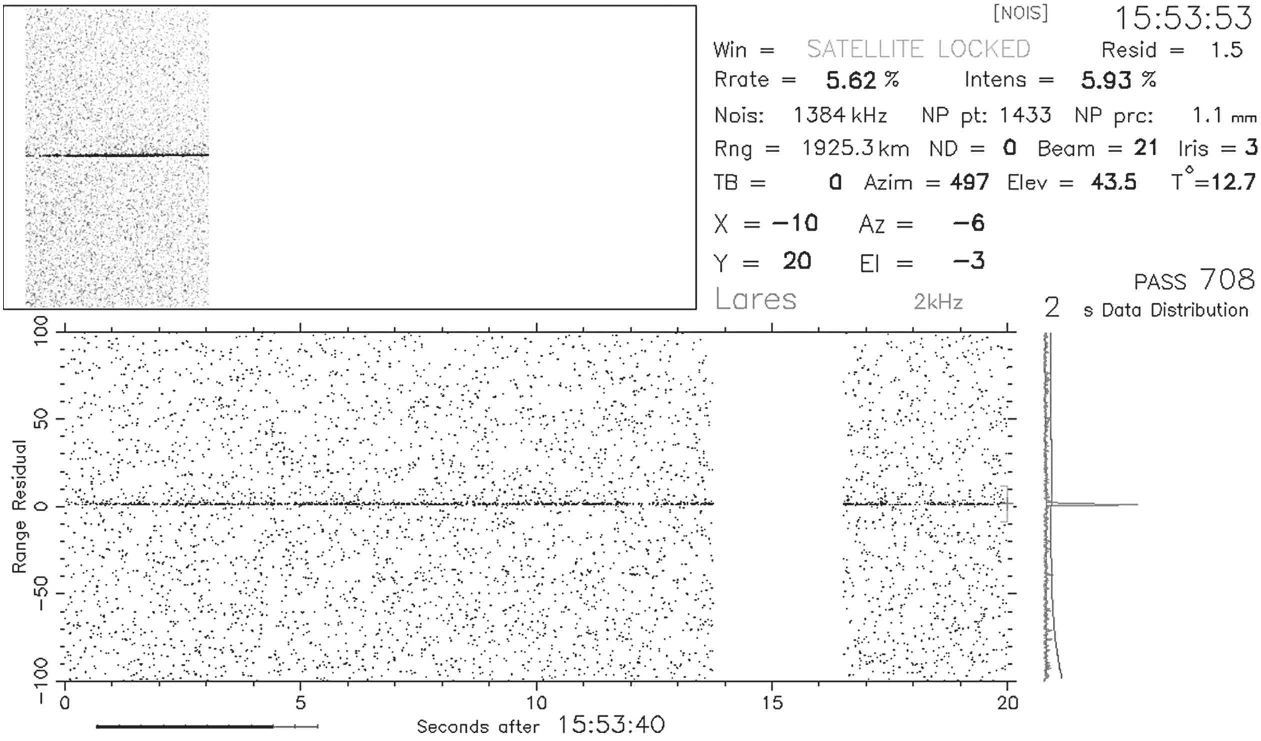
triggered the instantaneous predicted range in light-time is subtracted from the measured time-of-flight to form a range residual. Usually operators have access to real-time plotting of range residuals to aid tracking; an example of such a plot, from the UK Natural Environment Research Council system at Herstmonceux, is shown in Figure 12.4. Although noise points appear ubiquitous, actual satellite returns form a tight curve, the presence of which gives instant feedback to the observer (or to the closed-loop tracking system if present) that the target is in the beam and that range measurements are being achieved. The quality requirements for the orbit predictions are not too demanding, as stations have searching capabilities for tracking acquisition. However, if orbit predictions are not provided with sufficient frequency (usually a few days, depending on orbit height), their quality degrades and makes successful tracking a challenge.

12.3.1.2 Data Screening

Once the data have been obtained, it must be processed in order to discriminate between noise events and actual returns. Track-detection software, normally supervised by a human operator, is employed to make an initial coarse selection of true returns that inevitably includes some noise. Then either an orbit or a polynomial fit is computed to refine the selection. The polynomial fit method is simpler to compute, although there may be instances where it does not cope optimally with unusual residual data distributions or even show spurious oscillations. The orbit fit method performs similarly well, with the advantage that it provides a physical insight into the quality of the prediction, as it uses the data to compute corrections to the predicted orbital path of the satellite. The corrections usually take the form of small adjustments to the positions of the satellite in the across-track and along-track directions compared to those used a priori. Orbit corrections are determined iteratively, during which process noise points are removed by a rigorous procedure. The along-track corrections in particular can be useful to improve the accuracy of subsequent predictions.

12.3.2 Ranging Policy

Although the network of SLR stations is quite heterogeneous, we can make a distinction between essentially two different types of systems depending upon the original design specifications, which essentially dictate or impose significant constraints to their mode of operation. Some stations are designed to work at a multi-photon level of return, whereby each returning pulse of reflected laser light is typically at the level of at least a few photons, often measured at several hundreds of photons. Such systems use MCP detectors, which are very fast, compact, multiple photomultipliers in a single unit. The discriminating electronics associated with MCP detectors are usually configured such that they are triggered by the leading edge of the returning laser pulse, yielding a very precise range measurement, affected very little by the particular waveform of that pulse. All NASA systems of the previous generation fall into this category. The second type of system is designed to operate at or close to a single-photon level of *return*. The original philosophy behind the design of such systems was that the ability to detect such low return levels improves the chances of making observations under poor atmospheric conditions. For stations at near sea level in rapidly changing weather, typical of many European stations, this ability maintains their competitiveness with those systems working in drier conditions at high-altitude locations. The detectors commonly employed are SPADs, which are very small, solid-state devices with sensitive areas of the order of 100 μm in size. In order to ensure that the intensity of the return pulses does not exceed the single photon



```
                                    [NOIS]        15:53:53
Win =    SATELLITE LOCKED                  Resid =  1.5
Rrate =   5.62 %        Intens =   5.93 %
Nois:   1384 kHz   NP pt: 1433   NP prc:    1.1 mm
Rng =  1925.3 km  ND =  0  Beam = 21  Iris = 3
TB =        0  Azim = 497  Elev =  43.5   T°=12.7
X = −10     Az =    −6
Y =  20     El =    −3
                                               PASS 708
Lares              2kHz            2   s Data Distribution
```

FIGURE 12.4 Real-time display at SLR station of Herstmonceux, UK. Operators get instant feedback on signal acquisition, return rate, background noise levels, satellite pass timing details, and various relevant system settings.

level, it is necessary to monitor and adjust the detection rate in real time. Photon detection events follow Poisson statistics, which are used to estimate the average intensity from average return rates. For a typical detection quantum efficiency $q = 0.2$, provided the return rate is restricted to less than about 15% of laser shots fired, on average only about 0.6 photons reach the detector per shot, well within the single-photon regime. This may be verified as follows: the average number of photoelectrons that are detected is qn, where n is the number of photons impinging on the detector. For a detection to be made, clearly qn must be ≥ 1, so from Poisson statistics the probability of detecting a return may be expressed as

$$P\left(qn \geq 1\right) = 1 - P\left(qn = 0\right)$$

i.e.

$$P(qn \geq 1) = 1 - \frac{\mathrm{e}^{-(qn)}(qn)^0}{0!} = 1 - \mathrm{e}^{-(qn)}.$$

Which, expressed as a percentage return rate,

$$\text{rate} = \left(1 - \mathrm{e}^{-(qn)}\right) \times 100$$

The return rate may easily be computed in real time from the statistics of the last 50–100 shots fired, and the result used to control the neutral density wheel in real time.

Beyond the relative merits of these two kinds of systems in terms of their capabilities to yield successful returns at different meteorological conditions, a most fundamental disparity arises from the fact that they measure different optical reflection points in the satellite. Multi-photon systems strive to obtain high-energy returned pulses, thus ensuring that the photo-detector responds to light signals that originated from the closest retroreflectors facing the incident laser beam. The detector output in turn triggers the timing device in a consistent fashion, provided the required discrimination electronics and some degree of signal level control and calibration is in place. In contrast, the returns obtained by single-photon systems can originate from any point in the retroreflector array; the distribution of residuals they form is directly related, and *relatable*, to the geometry and characteristics of the target. In practise this requires slightly different approaches for the computation of the so-called centre of mass correction of the satellite, that is, the distance from the centre of mass of the object (the geometric point the equations of motion refer to) and the average optical reflection point. A decisive advantage of single-photon systems is that centre of mass corrections can be estimated without being subject to uncertainties about the behaviour of the detection electronics and the particular hardware settings in use (e.g. threshold and gain, use of additional amplifiers, time response of different devices). A perceived disadvantage of single-photon ranging is that the signal attenuation required to ensure a sufficiently low return intensity means that the data yield is in some circumstances lower than what could otherwise have been. This however only applies in cases where the return signals are higher than about 15% of the number of laser pulses fired, which is not always the case (and actually rare for high-altitude satellites for many systems). In any case, one of the key components in upgraded modern systems is the laser, with the overwhelming tendency to move towards higher firing rates (and lower energy per pulse) to shorten acquisition times. With this equipment, it is simply not feasible to operate at multi-photon levels of return. The growing consensus is that to achieve the ultimate accuracy the SLR technique can provide, single photon ranging is not an option, but a necessity.[1]

12.3.3 Calibration

In all SLR and LLR systems, whether permanent or transportable, it is necessary to refer the range measurements to an invariant reference point in the system. Most modern SLR telescopes are designed with altitude-azimuth mounts. The intersection of the azimuth and altitude axes, although generally not directly accessible, is fixed in the telescope structure and is defined as the station reference point. Using engineering drawings and arguments of rotational symmetry it is possible to define other nearby points which can be tied accurately into a local site survey. In particular, the distance from the telescope to one or more calibration targets can be measured to millimetre accuracy. Each time the laser is fired and a signal from the outgoing light pulse triggers the timer, that pulse has to travel, *via* a system of mirrors, up to the telescope before passing the reference point on its way to the satellite. Similarly, the return pulse passes the reference point before hitting the detector, the electrical pulse from which then travels to the timer. Thus, each time-of-flight measurement has added to it an extra time which, provided that the optical and electrical paths remain unchanged, is constant. Each range measurement must be corrected by the value of this constant so that it refers to the reference point. In actual fact there are residual temperature dependent effects and the value is frequently determined (e.g. hourly) by ranging to a ground target at a precisely known distance. It is also particularly important that the correction is thoroughly checked whenever any electrical, optical or mechanical alterations are made to any of the components involved in the hardware chain. For example changing the thickness of an optical filter or the length of a signal cable can induce changes in the total internal system delay. Efforts have been made at some stations to carry out calibration and ranging to satellites simultaneously, either by diverting a fraction of each outgoing pulse (*via* a known path in the telescope) or by devoting a small proportion of pulses to calibration. Neither of these approaches has been notably successful and dedicated ranging to ground targets, at a significant distance from the telescope or nearby inside the dome, is still very much in vogue.

12.3.4 Safety

Firing high-powered laser beams from ground level to space entails some safety risks that are addressed in multiple, redundant ways at the stations to ensure safe operations at all times. Laser irradiation to the operating personnel, as with every application involving the use of laser beams, is minimized by enforcing strict local safety regulations and procedures followed by adequately trained staff. Since the laser beams must of necessity be exposed at some point to the open environment, there are potential risks to unauthorized persons that must be taken into account. As most ranging stations are manned, this issue is simply addressed by prohibiting access to the facilities during operations. To minimize the risk of laser illumination of aircraft flying overhead several safety mechanisms are usually in place. First, although different aspects of scheduling, track search and a multitude of operational settings may be automated or greatly aided by computer control to several degrees, laser ranging still is a technique that requires considerable manual intervention. Thus, the presence of an operator able to stop the laser from firing in situations that may pose a risk is a primary safety line in most stations. In addition, usually a radar slaved to the emitter telescope is in place, which automatically shuts down the laser beam if an object is detected within a certain solid angle. This is an effective solution in most situations, although it may not be suitable in cases where electromagnetic interference is problematic for the performance of other co-located space geodetic techniques (e.g. VLBI). The use of ADS-B receivers and other transponder systems (e.g. FLARM) to obtain positional data from nearby aircraft is now common, with different levels of integration to the ranging systems, from purely informational to automatic laser locking. Visual and infrared cameras can also be employed to provide further information to a human operator, or in an automated fashion to inhibit firing. Lastly, some stations have agreements with the local aviation authorities to demarcate warning or exclusion zones for air traffic. The efficacy and redundancy of the various safety mechanisms in place at any given station mean that the risks are greatly minimized, as the decades of continuous routine operations of the global network attest.

12.3.5 Global Network

Beyond and above the level of individual stations, SLR activities are coordinated by the ILRS (Pearlman et al. 2002), which is one of the space geodetic services of the International Association of Geodesy (IAG), as well as a member of the IAG's Global Geodetic Observing System (Plag and Pearlman 2009). The role of the Services consists in providing the required geodetic infrastructure for monitoring global change in the Earth system. Standards and specifications proposals, tracking campaigns coordination, joint analysis efforts, data archiving, mission support approval, liaison with satellite operators, and international representation are some of the activities carried out by the ILRS. Currently there are over forty tracking stations active (shown in Figure 12.5), although the total number of sites that have at some point contributed laser ranging data is much greater. A historic weakness of the network has been its poor geographical distribution, with the bulk of the stations located in the northern hemisphere. This translates into a sub-optimal tracking coverage with wide gaps where no data are available, which has an impact on the quality of some geophysical parameters of interest derived from SLR measurements. In recent years this situation has improved somewhat, and it is expected that new planned stations will make a positive contribution in this

---

[1] The new Russian stations all operate in single photon, as the next generation of NASA systems will do.

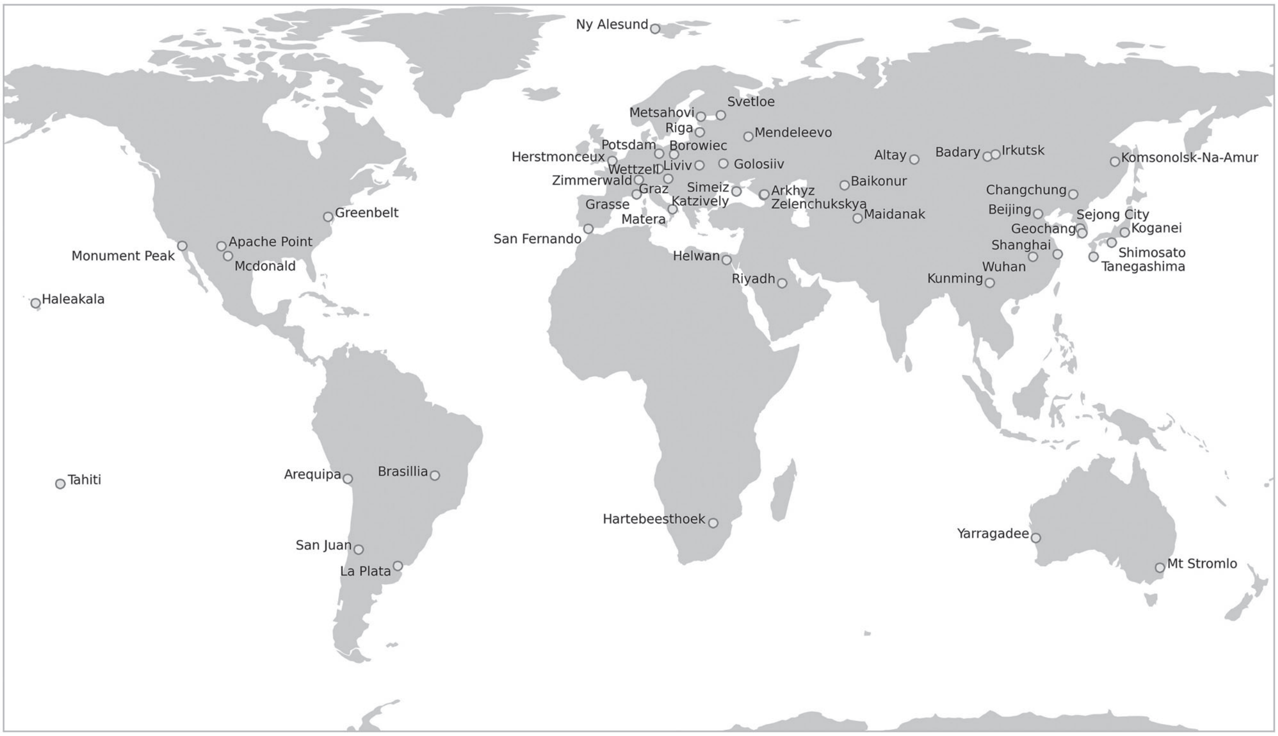

Ny Alesund
Svetloe
Metsahovi
Riga
Mendeleevo
Potsdam
Borowiec
Herstmonceux
Wettzell
Liviv
Golosiiv
Zimmerwald
Graz
Simeiz
Arkhyz
Zelenchukskya
Baikonur
Grasse
Katzively
Matera
San Fernando
Helwan
Riyadh
Maidanak
Altay
Badary
Irkutsk
Komsonolsk-Na-Amur
Changchung
Beijing
Geochang
Sejong City
Koganei
Shanghai
Wuhan
Kunming
Shimosato
Tanegashima
Greenbelt
Apache Point
Mcdonald
Monument Peak
Haleakala
Tahiti
Arequipa
Brasillia
San Juan
La Plata
Hartebeesthoek
Yarragadee
Mt Stromlo

FIGURE 12.5 Network of SLR stations of the ILRS. The number of stations in the northern hemisphere, particularly in Europe, is greater than in the south. Ideally, this level of densification would be available in other regions, allowing for an increased tracking yield and providing redundancy to counter data shortages due to poor local meteorological conditions. The heterogeneous geographical distribution of observing sites has an adverse effect on the quality of the global geodetic products.

respect in the future. At the moment the ILRS maintains a priority list of supported missions, on the basis of scientific value and specific tracking requirements. At times this is subject to alterations to provide intensive coverage during requested tracking campaigns, which sees the network concentrating its efforts on some specific objects. Otherwise there is no direct oversight of individual stations, which set up their priorities and organize their schedules according to their own capabilities, budgets, and any other, sometimes perhaps overlapping, activities.

12.4 Analysis

Satellite laser range observations obtained and pre-processed by the global network of tracking stations are promptly submitted to, and archived in, central facilities. Users can then access freely these data and use it for various purposes, usually in combined multi-satellite analyses employing observations taken over periods of several years. The ILRS through its Analysis and Associate Analysis Centres releases several analysis products on a continuous basis, and produces the combined solutions for its contribution to the realization of the terrestrial reference system. Most of the SLR analysis follows the same basic scheme, regardless of what investigation is being carried out: a mathematical model of the forces acting on the satellite is set up. An appeal is then made to Newton's law of universal gravitation, which states that the instantaneous force p acting on a body is proportional to the product of its mass m and its acceleration f, i.e. $p \sim mf$. Then, given an approximation to all the forces p, it is possible to calculate the instantaneous acceleration f, and by numerical integration determine the velocity and position of the satellite at a given instant. Once the path of the satellite about the centre of mass of the Earth has been estimated in this way, the expected range to the satellite from each observing station is computed at all the epochs for which SLR observations are available, using the best estimates of the instantaneous position of the ground stations. The computed ranges are then compared to the observed ones, and the set of differences used in a least-squares problem to solve for a whole host of corrections to the initial model parameters, the observing geometry, technique-dependent parameters, and possibly other geophysical quantities of interest. In this way, improvements are possible in the knowledge of, for example, the Earth's gravity field, ocean and solid Earth tides, atmospheric density at the heights of the satellites, and the effects of general relativity. Also possible, of course, is a very accurate determination of the paths of the satellites, which, especially for some remote-sensing satellites, is important in the analysis of their observations of, for example, the variations in heights of the oceans. As even the most cursory exposition of the mathematical details of orbit determination and parameter estimation would exceed the space constraints available in this piece, the interested reader is referred to Kaula (1966), Montenbruck and Gill (2000), and Beutler (2005) for excellent expositions to this topic.

12.4.1 Centre of Mass Corrections

In order to compare laser range measurements to the computed orbits, a correction must be made to take into account the distance offset of the reflection point from the centre of mass of the satellites, to which the orbits are referred. The value of this correction depends on the characteristics, size and physical distribution of the individual retroreflectors, the intensity of the return signal, the detection system employed at the station, and the observing policy in place. There are various approaches to compute the centre of mass corrections to retroreflectors in orbit, including the simulation of the far-field diffraction patterns (FFDP); convolution of theoretically or experimentally derived satellite impulse response functions; analysis of measured FFDP; and laboratory measurements using streak cameras or hardware set-ups similar to those employed during actual operations. The CoM corrections endorsed and made available by the ILRS for the geodetic satellites have been obtained from the computation of the impulse response functions for each target (Rodríguez et al. 2019). Briefly, the reflectivity of individual cube corners is computed for all possible incidence angles via ray-tracing; using the reflectivity and knowledge of the coordinates and orientation of each retroreflector, the response to an infinitesimally short incoming pulse is computed for a sufficiently large number of different orientations. The average response function is thus obtained and subsequently employed to model the returned signals for laser pulses of arbitrary length. Additionally, the effect of different intensity of returns can be taken into account from Poisson statistics considerations. A limitation of this approach is that knowledge of the response of the detection electronics (detectors, amplifiers, discriminators) is generally not available, and therefore their impact on the measured ranges of difficult quantification at accuracy levels approaching one millimetre. The accurate determination of centre of mass corrections as a function of ranging system parameters is of the maximum importance, and still the subject of active research both for new and long-established targets (Otsubo et al. 2014, Kucharski et al. 2015; Rodríguez et al. 2019).

12.4.2 Accuracy and Quality Control

Multiple metrics are used to assess different aspects of the performance of individual stations. Various measures of productivity, precision, short and long-term stability, pass and pass segment coverage are common ones. In terms of data yield, there are considerable differences between stations of the network, as the primary obstacle to laser ranging operations is weather conditions. SLR requires sufficiently clear skies for the laser pulses to reach the satellites and be reflected back. Depending on the target, tracking operations are possible under fairly poor atmospheric conditions (horizontal visibilities lower than 5 km) and even a moderate presence of cirrus, but successful laser returns cannot be obtained through any sort of dense clouds. Hence, local weather conditions put a ceiling on the productivity of laser stations. The precision achievable by the SLR technique has been historically limited by the length of the laser pulses and the jitter of the timing devices. Nowadays the short pulse lengths (down to 10 ps in modern lasers), timing precision of modern event timers (a few ps), and high firing rates (up to multi-kHz) make it possible to achieve precision levels theoretically below 1 mm. These advances reveal sources of variability whose contributions had hitherto been not deemed significant. For instance, the fast data acquisition and improved precision of modern kHz systems allow to obtain returns from individual retroreflectors, which are normally not visible in previous SLR generations; opening the possibility for new data reduction techniques to further improve measurement precision. It must be noted that for the most part, most users of laser ranging data employ the so-called normal point (NP) data instead of raw ranges. NPs are a concise representation of the ranging measurements computed by averaging the residuals to a smoothing function over fixed periods of time. The duration of NPs is satellite specific (related to orbit height) and is chosen ensuring that no useful information is lost over the raw data. The NP precision is of course greater thanks to the accumulation of n single ranging measurements, improving as root(n).

The stability of tracking stations over long periods of time is of paramount importance, as well as a challenge. The requirement is no less than ensuring consistent performance from each hardware sub-system, under changing environmental conditions, for a variety of different targets, operating a system that must be continuously monitored for faults and unexpected behaviour. Stability is assessed employing data from the global network to derive orbit solutions against which the ranging data of any given station can be compared. This is the basis of the various short-term quality control systems in place that monitor the performance of the network. Assessments of absolute accuracy are possible by analyses of several years of data in orbital dynamics solutions that allow the explicit estimation of system error parameters. Results obtained this way reveal that the best performing stations of the network consistently achieve absolute accuracies at the level of less than 5 mm over many years (Appleby et al. 2016). Crucially, this kind of analyses cannot make a distinction between different sources of error, and therefore can only point to the existence of a problem rather than its origin, which could lie anywhere from the measuring equipment and its operation to the modelling assumptions and standards employed.

12.5 Applications

SLR data are used for several purposes, from the most theoretical aspects, such as in experiments aiming to test certain relativistic effects, to more practical ones, such as in orbit validation studies for satellite missions equipped with their own navigation systems. Some of the most notable applications of SLR follow below.

12.5.1 Terrestrial Reference Frame

A practical product of space geodesy of primary importance is the realization and maintenance of a high-precision, global reference frame. In essence this means determining the geocentric coordinates and velocities of a worldwide set of observing stations, as well as the time-series of parameters that describe the instantaneous orientation of the Earth. The Earth system is subject to a whole range of forces that cause it to deform, wobble, and rotate in not a strictly regular fashion as it moves through space. Because the Earth is an oblate spheroid, with polar diameter some 50 km less than the equatorial diameter, the gravitational attractions of the Moon and Sun on the Earth cause its axis both to process slowly, completing one cycle every 26 000 years, and undergo a series of short-period nutations, or wobbles, on timescales of tens of years. Tides raised by the Moon and Sun, in both the oceans and in the solid Earth, cause periodic displacements of several tens of centimetres at the surface, depending upon location. In addition, since the whole Earth system comprises atmosphere, solid earth, oceans, and liquid core, there is angular momentum exchange between these elements causing seasonal, short-, and long-period changes in the rotation rate of the solid Earth, and polar motion effects. Plate tectonic motions may be caused by the production of new crust as molten rock cools on reaching the surface or by crust loss as one plate subsides beneath another. The result is that sites on the surface of the Earth may be slowly drifting apart or coming together at rates of up to 5 cm year^{-1}. For these reasons, in order to define the instantaneous position of an observer on the surface of the Earth, a reference system must be set up comprising a set of coordinates and velocities of fiducial stations, and a series of theoretical and empirical parameters describing the various aspects of Earth rotation, to relate that frame to an inertial frame reference. Such a system is required, e.g. to link in space and time scientific observations taken on or above the surface of the planet; to align to a common global reference system regional surveys and mapping; or to relate with high accuracy the position of tracking stations to those of spacecraft in orbit in the solar system so that efficient communication links can be established.

Four space geodesy techniques contribute their various strengths to realize and maintain the global reference system. The orientation of the Earth in space is addressed by VLBI. Accurate VLBI site coordinates are determined relative to an extra-galactic set of radio sources, simultaneously with all the components of Earth's rotation. Global navigation satellites measurements using a ground network of hundreds of receivers on the surface of the earth are used to derive horizontal velocity fields and thus define in detail the relative motions of tectonic plates. In addition, GNSS is the technique employed to distribute the reference frame to all kinds of users, as receivers are inexpensive, light weight, and nowadays ubiquitous. Doppler orbitography and radiopositioning integrated by satellite (DORIS), an independent orbit determination and positioning system, contributes station coordinates and Earth orientation parameters. The SLR technique itself is critically important in two major areas of the maintenance of the reference frame, namely, its scale and origin. Given perfect instrumentation, the accuracy of laser ranging measurements does not depend upon orbit height: the distance to the Moon can be measured as accurately as can the distance to a low-Earth orbiting geodetic satellite. As the paths of the satellites themselves are, of course, fundamentally sensitive to the location of the geocentre, the sizes of the orbits can be determined accurately; this in turn leads to an accurate determinations of the scale (Dunn et al. 1999) and origin (Spatar et al. 2015) of the reference frame, from the locations of the tracking stations.

The most recent determination of the ITRF, a combined solution from VLBI, SLR, GPS, and DORIS, is ITRF2014 (Altamimi et al. 2016). ITRF2014 provides coordinates and velocities of several hundred geodetic stations, all referred to the same consistent reference frame. The SLR contribution consisted of seven independent solutions from the ILRS Analysis Centres.

12.5.2 Support of Scientific Missions

Many satellite missions involved in a multitude of scientific applications have been and are currently tracked by SLR. When the scientific goals demand the utmost level of orbit accuracy and precision, the inclusion of an onboard laser retroreflector array enables laser ranging operations in addition to other tracking systems. Among the missions that have in the past required SLR support we find an array of scientific goals and objectives, such as global environmental and disaster monitoring; high-resolution observation of the earth's surface; geomagnetism studies; measuring changes in the cryosphere; mapping of ocean surfaces and exploration of ocean circulation; measurement of Earth's gravity field; testing of theories of fundamental physics; technology demonstration and testing; calibration of onboard navigation equipment; and a wide range of Earth observation satellites with different instrumental payloads.[2]

A particularly notable subset of scientific missions supported by SLR ranging are altimetry satellites, which make range measurements from their orbits to ocean, ice, and, in some cases, land surfaces. These measurements are used for research into ocean currents, sea-level monitoring, ice mass-balance studies among others. In order to process the altimetry results, and in particular refer them to a global, consistent reference frame with an origin at the centre of mass of the Earth, the orbit of the satellite must be determined within the same reference frame and to the same level of accuracy. The strengths of SLR as a tracking system in these missions reside in providing an absolute validation of the radial accuracy of their orbits, and depending on the analysis strategy, contributing directly to the dynamic orbit determination along with other navigation data (GNSS, DORIS). The accuracy and stability requirements that SLR must meet are particularly stringent, in order to be able to, for example, identify geographically correlated orbit errors at inter-annual and decadal time scales at the 1 mm/y level (Couhert et al. 2015).

SLR and LLR have proven to be a useful tool for testing some experimental consequences of general relativity. Geodetic satellites, with their small area-to-mass ratios to minimize non-gravitational perturbations, provide a close realization of the concept of a test mass. Studying the motion of these satellites, by comparing the experimental, measured orbits to the predicted ones computed from a particular set of dynamic models, provides a simple test of gravitational dynamics (Ciufolini et al. 2011, Peron 2014, Lucchesi et al. 2015).

---

[2] https://ilrs.cddis.eosdis.nasa.gov/missions/satellite_missions/past_missions/index.html

This is possible thanks both to the intrinsic precision of laser ranging data and the accurate modelling of satellite motion, which must take into account all possible forces acting on the objects under consideration to be able to extract the signal of the small relativistic effects. Measurements of the secular relativistic precession produced by the Lense-Thirring effect were first carried out using laser ranging data of the LAGEOS satellites. Subsequent studies, taking advantage of better geopotential models, would see improvements in precision and accuracy (Ciufolini et al. 1998, Ciufolini and Pavlis 2004). The latest addition to the constellation of geodetic satellites, the Italian Space Agency mission LARES (Laser Relativity Satellite), has the primary scientific objective of measuring the Earth's dragging of inertial frames predicted by general relativity, for which results have already been obtained using a combination of both LAGEOS and LARES data (Ciufolini et al. 2016; Lucchesi et al. 2017). Other effects predicted by general relativity have been measured using SLR data. The relativistic precession of the argument of pericentre has been measured using observations to the LAGEOS-2 satellite and found to be in agreement with GR predictions to a level of 99.8% (Lucchesi and Peron 2010). An experiment to test the gravitational redshift predicted by the Einstein equivalence principle is currently ongoing, employing the onboard clocks of two Galileo spacecraft that due to a technical problem were inserted in elliptical orbits (Delva et al. 2015). The ILRS has organized an intensive laser ranging campaign to support this experiment, which will use SLR data to dis-entangle clock and orbit errors.

12.5.3 Tracking of GNSS Constellations

The ILRS network makes regular observations of several GNSS constellations and organizes intensive tracking campaigns when required. The orbits of these satellites, the so-called 'micro-wave orbits', are determined operationally from the navigational radiometric data. SLR tracking data are employed to validate the micro-wave orbits (Urschl et al. 2007, Sośnica et al. 2017); for orbit improvement purposes (Hackel et al. 2015); for the determination of certain technique-specific parameters (satellite antenna and laser retroreflector offsets) and improvement of dynamic models (Montenbruck et al. 2015). Additionally, GNSS and SLR data can be combined for the determination of geodetic parameters (Thaller et al. 2011, Thaller 2014). The comparison/validation process of GNSS orbits may be carried out in two distinct ways: first, by computing independent orbits using SLR data alone and comparing them with the micro-wave orbits positions; or directly by comparing the laser range measurements with the corresponding distances derived from the micro-wave orbits. A direct comparison between micro-wave orbits and SLR measurements has the advantage that the comparison is not degraded by potential systematic effects in the orbits that are determined using the SLR data alone, which can be quite sparse. The method does, however, only provide a check on the range of the satellites as computed from the micro-wave orbits. Although a range error for these high-orbiting satellites will be dominated by an orbital radial error, some contamination from, e.g. along-track errors, is inevitable.

12.5.4 Space Debris

The near Earth space environment has been increasingly populated by human-made objects since the first days of the space age in the mid-20th century. Space debris are non-functional objects orbiting the Earth, which include large, no longer operational satellites and rocket bodies, as well as small fragments created by collisions between larger objects, and even smaller particles with sizes below a millimetre. Naturally, space debris occupies the same space as active missions of all kinds and therefore represents a significant threat to the integrity of costly equipment and therefore to the myriad of services they provide. Moreover, the increasing numbers of space debris may jeopardize the use the space environment by future generations, and it has been recognized as a problem of sustainable use of resources (Schildknecht 2007). Efforts to monitor and study the population and evolution of space debris have been initiated by several countries, employing for the observational tasks radar and optical techniques.

The adequacy and feasibility of laser observations to tackle certain aspects related to space debris is a topic only explored in recent years. Laser ranging measurements to both defunct satellites carrying retroreflectors as well as to uncooperative targets (e.g. spent upper stages) can be employed to improve the orbit determination of these objects; this in turn may be invaluable in situations of potential collision with other operational missions in order to schedule, and assess the necessity of, costly avoidance manoeuvres (Bennett et al. 2013). Technical progress in this field has already been demonstrated, with several groups operating, or planning to in the near future, space debris-capable laser ranging systems (Zhang et al. 2012, Kirchner et al. 2013). Tracking defunct satellites has enabled the determination of their attitude and spin parameters (Kucharski et al. 2014) as well as an accurate estimation of the direct solar radiation pressure force acting on the object (Kucharski et al. 2017). It is expected that more than twenty stations worldwide, between new and upgraded systems, will be able to track uncooperative targets in the following years.

12.5.5 Time Transfer

Time transfer is the comparison of clocks located in remote locations. The most widely used techniques for time transfer are Global Positioning System (GPS) and Two-Way Satellite Time and Frequency Transfer. The use of laser links for time transfer had been demonstrated at inter-continental distances with the LASSO (Laser Synchronization from Stationary Orbit) experiment in the early 1990s, achieving sub-ns precisions (Fridelance and Veillet 1995). About two decades later, the Time Transfer by Laser Link (T2L2) experiment was initiated, with the in-orbit payload hosted by the Jason-2 satellite (Fridelance et al. 1997, Samain et al. 2008, Samain 2015). T2L2 exploits the existing network of laser stations, which perform accurately time-tagged tracking of the retroreflector onboard Jason-2. The laser pulses are also detected and recorded by the space equipment, allowing the subsequent comparison of the ultra-stable ground clocks. The design goals included a time stability in the range of 1 ps over 1000 s and an accuracy better than 100 ps (Exertier et al. 2014). Calibration of ground stations and intensive tracking campaigns were carried out to support T2L2, which demonstrated the possibility to independently validate other time-transfer techniques, leading to consistencies on continental links below 250 ps (Exertier et al. 2016). In the most recent campaign, accurate modelling of the onboard oscillator enabled to perform time transfer in a non-common view mode over inter-continental distances, achieving a consistency between GPS and the laser link technique below 1 ns, and sub-100 ps between participating European stations (Samain et al. 2017).

Successful time transfer using SLR systems has been conducted between the Chinese Beidou navigation satellites and laser ground stations (Meng et al. 2013). A similar experiment under development is the European Laser Timing experiment, which aims to establish a link between the ground clocks and the timescale provided by atomic clocks onboard the International Space Station with an accuracy of 100 ps (Schreiber et al. 2009).

12.5.6 Lunar Laser Ranging

During the 1960s and early 1970s Russian and US missions deployed retroreflector arrays on the surface of the Moon in order that LLR might be attempted. Currently five arrays, three American units placed during the Apollo missions, and two French ones loaded on the Soviet rovers Lunokhod are routinely employed by the few lunar-capable stations of the network. As the reduction in the link budget is proportional to the fourth power of the distance, ranging to the Moon is more than 150 000 times more difficult than to GNSS satellites, making this kind of measurements technically very challenging. Only five laser stations are able to range to the Moon, of which just three perform regular measurements. Critical parameters for successful LLR are high laser power; minimal beam divergence; accurate pointing and tracking; big telescope aperture (Murphy 2013).

LLR data are employed for general relativity testing, e.g. tests of the Equivalence Principle, time variation of the gravitational constant, and the inverse square law (Williams et al. 2004, Müller et al. 2007, Merkowitz 2010, Williams et al. 2012). The availability of ranging measurements has improved, relative to classic optical data, the accuracy of the lunar orbit determination by orders of magnitude. LLR has been extensively used for geodynamics research, eliciting studies on the long-term variations of Earth's rotation, precession and nutation, gravity field and tides. In addition, information about the internal composition and dynamics of the moon is obtained from LLR data, such as low-degree geopotential harmonics coefficients, moments of inertia, lunar Love number, and the inference of the presence of a liquid core in the Moon (Dickey et al. 1994, Murphy 2013).

12.6 Concluding Remarks and Future Trends

Understanding the processes that drive global change in the Earth system is a requirement for modern societies, in order to adopt informed policy decisions to minimize the negative impact some aspects of those changes may exert; to adapt in the long term to the environmental realities of a changing planet; and to plan and coordinate in advance at both national and international levels prevention strategies and sustainable development models to ensure the well-being of future generations. Geodesy is the discipline that at a fundamental level underpins the scientific efforts aimed at gaining that understanding. By providing long-term monitoring of the solid and liquid Earth, temporal variations in the gravity field, the rotation of the planet, and a global reference frame upon which consistent observations of the Earth system can be realized, geodesy furnishes geosciences with a foundational framework essential for the study and interpretation of global change processes. The recognition of the role of geodesy, and in particular the realization of a global reference frame, was recognized by the United Nations resolution on Global Geodetic Reference Frame for Sustainable Development.[3] SLR is and will remain a key technology for these continuing endeavours. Increasingly stringent requirements in terms of precision, accuracy, and stability mean that constant improvements must be realized in the observing networks of the various techniques involved, the data processing strategies, and models employed in the data analysis. To this end many stations of the laser ranging network are upgrading their systems, and new sites are either under construction or at the planning stage. In addition, base technologies upon which key components of SLR systems rely continue to evolve, driving improvements in the technique and enabling new applications. Examples of this include the ultra-short pulse length, multi-kHz firing rate lasers (e.g. up to 100 kHz) that are currently available in form factors that make possible new, simpler system configurations; improved detection technologies with high efficiency in the infrared region; compact, inexpensive event timer devices; and a general tendency towards lower costs and widespread availability of most hardware components. The uptake of new technology as it becomes available translates directly into improvements in measurement quality, which in turn drives improvements in the data processing and refinement of the background models, ultimately resulting in superior products. Since its origins, SLR has seen remarkable improvements, evolving on the back of technological progress, becoming the mature space geodetic technique it is today. If the objective in the 1970s and early 1980s could be said to have been the invention and first refinements of the technique, that of the 1990s the consolidation of the network and emergence of new sites, and with the latest significant modernization of the hardware taking place in the late 2000s, the final aim for SLR in the present time and in the near future is the quest for ultimate performance reaching the goal of 1 mm level accuracy.

REFERENCES

Altamimi Z., Rebischung P., Métivier P., Collilieux X. (2016) ITRF2014: a new release of the International Terrestrial Reference Frame modeling nonlinear station motions. *J. Geophys. Res. Solid Earth* 121(8):2169–9356.

---

[3] http://ggim.un.org/knowledgebase/KnowledgebaseCategory37.aspx

Appleby G.M., Rodriguez J.C., Altamimi Z. (2016) Assessment of the accuracy of global geodetic satellite laser ranging observations and estimated impact on ITRF scale: estimation of systematic errors in LAGEOS observations 1993–2014. *J. Geod*. DOI:10.1007/s00190-016-0929-2.

Bennett J.C., Sang J., Smith C.H., Zhang K. (2013) Accurate orbit predictions for debris orbit manoeuvre using ground-based lasers. *Adv. Space Res.* 52(11):1876–1887.

Beutler G. (2005) *Methods of Celestial Mechanics* (Berlin: Springer).

Bosworth J.M., Coates R.J., Fishetti T.L. (1993) The development of NASA's crustal dynamics project. Contrib. Space Geodesy to Geodynamics: Technol. Geodynam. Series 25:1–20.

Ciufolini I. et al. (2011) Testing gravitational physics with satellite laser ranging. *Eur. Phys. J. Plus* 126:72.

Ciufolini I. et al. (2016) A test of general relativity using the LARES and LAGEOS satellites and a GRACE Earth gravity model. *Eur. Phys. J. C*. 76:120.

Ciufolini I., Pavlis E.C, Chieppa F, Fernandes-Vieira E., Perez-Mercader J. (1998) Test of general relativity and measurement of the Lense-Thirring effect with two Earth satellites. *Science* 279:2100–2109.

Ciufolini I., Pavlis E.C. (2004) A confirmation of the general relativistic prediction of the Lense-Thirring effect. *Nature* 431:958–960.

Couhert A. et al. (2015) Towards the 1 mm/y stability of the radial orbit error at regional scales. *Adv. Space Res.* 55(1):2–23.

Courde C. et al. (2017) Lunar laser ranging in infrared at the Grasse laser station. *Astron. Astrophys.* 602:A90.

Degnan J. (1985) Satellite laser ranging: current status and future prospects. *IEEE Trans. Geosci. Remote Sens.* GE-23:4.

Degnan J.J. (1993) Millimeter accuracy satellite laser ranging: a review. Contrib. Space Geodesy to Geodynamics: Technol. Geodynam. Series 25:133–162.

Delva P., Hees A., Bertone S., Richard E., Wolf P. (2015) Test of the gravitational redshift with stable clocks in eccentric orbits: application to Galileo satellites 5 and 6. *Class. Quantum Grav.* 32 232003

Dickey J.O. et al. (1994) Lunar laser ranging: a continuing legacy of the Apollo program. *Science* 265:482–490

Dunn P. et al. (1999) Earth scale defined by modern satellite ranging observations Geophys. *Res. Lett.* **26** 1489–1492.

Exertier P. et al. (2014) Time transfer by laser link: data analysis and validation to the ps level. *Adv. Space Res.* 54(11):2371–2385.

Exertier P. et al. (2016) Sub-ns time transfer consistency: a direct comparison between GPS CV and T2L2. *Metrologia* 53:1395–1401.

Fridelance P. Samain E., Veillet C. (1997) T2L2 – time transfer by laser link: a new optical time transfer generation. *Exp. Astron.* 7(3):191–207.

Fridelance P., Veillet C. (1995) Operation and data analysis in the LASSO experiment. *Metrologia* 32:27–33.

Hackel S., Steigenberger P., Hugentobler U., Uhlemann M., Montenbruck O. (2015) Galileo orbit determination using combined GNSS and SLR observations. *GPS Solut.* 19:15–25.

Kaula W.M. (1966) *Theory of Satellite Geodesy: Applications of Satellites to Geodesy* (New York: Dover Publications).

Kirchner G. Koidl F., Friederich F., Buske I., Völker U., Riede W. (2013) Laser measurements to space debris from Graz SLR station. *Adv. Space Res.* 51(1):21–24.

Kucharski D. et al. (2014) Attitude and spin period of space debris Envisat measured by satellite laser ranging. *IEEE Trans. Geosci. Remote Sens.* 52(12):7651–7657.

Kucharski D. et al. (2017) Photon pressure force on space debris TOPEX/Poseidon measured by satellite laser ranging. *Earth and Space Science*. DOI:10.1002/2017EA000329.

Kucharski D., Kirchner G., Otsubo T., Koidl F. (2015) A method to calculate zero-signature satellite laser ranging normal points for millimeter geodesy – a case study with Ajisai. *Earth, Planets and Space* 67:34.

Lightstone A.W. and McIntyre R.J. (1981) Avalanche photodiodes for single photon detection. *IEEE Trans. Electron Dev.* ED-28:1210.

Lucchesi D.M. et al. (2015) Testing the gravitational interaction in the field of the Earth via satellite laser ranging and the Laser Ranged Satellites Experiment (LARESE). *Class. Quantum Grav.* 32. DOI:10.1088/0264-9381/32/15/155012.

Lucchesi D.M., Magnafico C., Peron R., Visco M. (2017) The LARASE research progam. State of the art on modelling and measurements of general relativity effects in the field of the Earth: a preliminary measurement of the Lense-Thirring effect. *MetroAeroSpace. IEEE International Workshop*. DOI:10.1109/MetroAeroSpace.2017.7999552.

Lucchesi D.M., Peron R. (2010). Accurate measurement in the field of the Earth of the general relativistic precession of the LAGEOS II pericenter and new constraints on non-Newtonian gravity. *Phys. Rev. Lett.* 105:231103.

Meng W. et al. (2013) Design and experiment of onboard laser time transfer in Chinese Beidou navigation satellites. *Adv. Space Res.* 51(6): 951–958.

Merkowitz S.M. (2010) Tests of gravity using lunar laser ranging. *Living Rev. Relativ.* 13:7.

Montenbruck O. and Gill E. (2000) *Satellite Orbits—Models, Methods and Applications* (Berlin: Springer).

Montenbruck O., Steigenberger P., Hugentobler U. (2015) Enhanced solar radiation pressure modeling for Galileo satellites. *J Geod.* 89:283–297.

Müller J., Williams J.G., Turyshev S.G., Shelus P.J. (2007) Potential capabilities of lunar laser ranging for geodesy and relativity. In *Dynamic Planet: Monitoring and Understanding a Dynamic Planet with Geodetic and Oceanographic Tools*. IAG Symposium Cairns, Australia. Tregoning P. ed. Springer Berlin Heidelberg, Berlin, pp. 903–909.

Murphy T.W. (2013) Lunar laser ranging: the millimeter challenge. *Rep. Prog. Phys.* 76:076901.

Otsubo T. and Appleby G.M. (2003) System-dependent center-of-mass correction for spherical geodetic satellites. *J. Geophys. Res. Solid Earth* 108(B4):2201. DOI:10.1029/2002JB002209.

Otsubo T., Sherwood R.A., Appleby G.M., Neubert R. (2014) Center-of-mass corrections for sub-cm-precision laser-ranging targets: Starlette, Stella and LARES. *J. Geod.* DOI:10.1007/s00190-014-0776-y.

Pearlman M.R., Degnan J.J. and Bosworth J.M. (2002) The international laser ranging service. *Adv. Space Res.* **30**:135–143.

Peron R. (2014) Testing general relativistic predictions with the LAGEOS satellites. *Adv. High Energy Phys.* 2014:12. DOI:10.1155/2014/791367.

Plag H-P., Pearlman M. (eds.) (2009) *Global Geodetic Observing System: Meeting the Requirements of a Global Society on a Changing Planet in 2020* (Springer). ISBN 978-3-642-02687-4.

Plotkin H. (1964) S66 Laser Satellite Tracking Experiment. Quantum Electronics III, vol. 2, Columbia University Press, New York, pp. 1319.

Rodríguez J., Appleby G., Otsubo T. (2019). Upgraded modelling for the determinatino of centre of mass corrections of geodetic SLR satellites: impact on key parameters of the terrestrial reference frame. *J. Geod.* 93:2553-2568. doi:10.1007/s00190-019-01315-0.

Samain E. (2015) Clock comparison based on laser ranging technologies. *Int. J. Modern Phys. D* 24(8): 1530021.

Samain E. et al. (2008) Time transfer by laser link – the T2L2 experiment on Jason-2 and further experiments. *Int. J. Modern Phys. D* 178(7):1043–1054.

Samain E. et al. (2017). Time transfer by laser link (T2L2) in non-common view between Europe and China (in press).

Schildknecht T. (2007) Optical surveys for space debris. *Astron. Astrophys. Rev.* 14:41–111.

Schreiber U., Prochazka I., Hugentobler U., Schäfer W., Cacciapuoti L, Nasca R. (2009) The European Laser Timing (ELT) experiment on-board ACES. *Frequency Control Symposium. IEEE International*. DOI:10.1109/FREQ.2009.5158251.

Sośnica K. et al. (2017) Validation of Galileo orbits using SLR with a focus on satellites launched into incorrect orbital planes. *J. Geod.* DOI:10.1007/s00190-017-1050-x.

Spatar C.B., Moore P., Clarke P.J. (2015) Collinearity assessment of geocentre coordinates derived from multi-satellite SLR data. *J. Geod.* 89:1197–1216.

Thaller D, Sośnica K, Dach R, Jäggi A, Mareyen M, Richter B, Beutler G (2014) Geocenter coordinates from GNSS and combined GNSS–SLR solutions using satellite co-locations. In: Rizos Ch and Willis P (eds), *Earth on the Edge: Science for a Sustainable Planet,* vol 139, International Association of Geodesy Symposia, pp 129–134, doi10.1007/978-3-642-37222-3_16.

Thaller D., Dach R., Seitz M., Beutler G., Mareyen M., Richter B. (2011) Combination of GNSS and SLR observations using satellite co-locations. *J. Geod.* 85(5):257–272.

Urschl C., Beutler G., Gurtner W., Hugentobler U., Schaer S. (2007) Contribution of SLR tracking data to GNSS orbit determination. *Adv. Space Res.* 39(10):1515–1523

Williams J.G., Turyshev S.G., Boggs D.H. (2004) Progress in lunar laser ranging tests of relativistic gravity. *Phy. Rev. Lett.* 93:261101.

Williams J.G., Turyshev S.G., Boggs D.H. (2012) Lunar laser ranging tests of the equivalence principle. *Class. Quant. Grav.* 29 184004

Zhang Z-P., et al. (2012) The use of laser ranging to measure space debris. *Res. Astron. Astrophys.* 12(2):212–218.

FURTHER READING

Other introductory chapters on the technique of satellite laser ranging, with a different emphasis, are available. These are contained within texts that cover the wider field of space geodesy in an excellent manner:

Combrinck L. (2010). Satellite laser ranging. In Sciences of Geodesy – I: Advances and Future Directions. Xu, G. ed. Springer Berlin Heidelberg, Berlin. ISBN: 978-3-642-11741-1.

Seeber G. (2003). Satellite Geodesy: Foundations, Methods, and Applications. 2nd ed. (Walter de Gruyter, Berlin, New York). ISBN: 3-11-017549-5. Ch. 8, pp. 404–442.

13

Lasers in Medical: Section Introduction

Terence A. King and Brian C. Wilson

The remarkable properties of laser radiation give the laser a rich diversity of applications in medicine and surgery. Laser radiation can be closely controlled, focused and manipulated to give precise, specific and localized effects in tissues. When used with flexible optical fibres, convenient delivery of the radiation is possible and internal organs of the body may be treated. There are many available laser wavelengths which can be selected to match tissue absorption. The characteristics of radiation, such as power and energy density, can be varied over large values. Lasers are now used in almost all medical or surgical specialties for the treatment of many diseases. In addition to therapeutic applications, there has developed a wide range of applications in biomedical science, diagnostics, imaging and sensing.

The evolution of the subject in these areas reflects, to a large extent, the synergism between physics and engineering, and biology and clinical applications. The use of the laser has stimulated the development of new techniques in medicine. Associated with this is the fact that the laser and optical industries have a key role to play in bringing forward new and improved sources and technologies. Medical applications have often benefited from the advances in technology originally developed for other purposes. Biomedical applications have been able to call on many advanced laser sources, optical techniques and components, particularly optical fibre techniques, which have been developed for other primary applications in engineering and telecommunications. It has also inspired the development of new techniques, such as optical coherence tomography for medical imaging.

The wide range of laser sources is described in detail in this Handbook. Laser technology is evolving and new lasers continue to be developed which provide new wavelengths or improved radiation characteristics which have potential applications in medicine and biology. The commercial sources for medical application generally are becoming more refined and compact. Semiconductor laser diodes are being further implemented, both for using directly or to excite crystal solid-state and fibre lasers with improved radiation characteristics for medical application. New optical imaging and sensing techniques extend the way in which structural and functional information in biomedical systems may be measured and analyzed. There is a great versatility in laser radiation which can be manipulated in many ways to extend their characteristics. This adds to the range and effectiveness in medical application.

The range of laser wavelengths available for medical application extends from the near-ultraviolet to the mid-infrared, and more recently, femtosecond lasers have been used for the generation of terahertz radiation with the potential for medical application. The wavelength range can be further extended using non-linear optical wavelength conversion techniques such as harmonic generation, sum and difference frequency mixing and optical parametric oscillators (OPOs). Power densities over many orders of magnitude are achievable with lasers, either directly or resulting from the spatial coherence of the laser beam, enabling very small focused spot sizes. The ability of lasers to produce pulsed output with pulse durations from milliseconds to femtoseconds provides high peak power for tissue interactions. Ultrashort pulse lasers also have growing diagnostic applications, for example, in multi-photon microscopy and fluorescence lifetime imaging.

There has been a steady development in the understanding of the mechanisms of the interactions of laser radiation with bodily tissues. The fundamental interaction processes are reviewed in Chapter D5.1 by Jacques and Patterson; this review provides an overview of the underlying processes for diagnostic and therapeutic applications. The development of an understanding of the laser–tissue interactions and the biological effects in living tissue is a vital step in determining optimum laser treatment. As a low-invasive tool, the laser is nowadays widely used in ophthalmic laser therapy and surgery (Chapter D5.2), refractive surgery (Chapter D5.3) and photodynamic therapy (Chapter D5.4). Some of the key therapeutic applications are described in Chapters D5.5–D5.7, which include tumour therapy, ophthalmology and photorefractive surgery, photodynamic therapy, dentistry, dermatology and vascular surgery. In Chapter D5.8 tissue interaction studies with the free electron laser are reviewed by Jansen et al. The wide wavelength range of this laser, its tunability and pulse structure enables tissue interactions to be investigated in new wavelength regions, so that new sources for specific applications can be identified.

In recent years there have been substantial advances in the use of lasers in biomedical science, diagnostics and imaging techniques, particularly with the development of photon time-of-flight measurements, confocal scanning, multi-photon imaging and fluorescence- and Raman-based techniques. Many of these techniques can be used in patients for diagnosis or to guide and monitor treatments, as reviewed in Chapter D5.9 by Wilson. Chapter D5.10 is devoted to broad bandwidth light sources in optical coherence tomography. In addition, many of these techniques, plus new developments such as optical trapping and tweezers giving laser manipulation of cells and biomolecules, surface plasmon resonance can be applied to address biological questions. The application of laser and optical techniques in biology and biotechnology is developing apace and this field is reviewed by Wachsmann-Hogiu, Annala and Farkas in Chapter D5.11. In the application of lasers in medicine particular issues on laser safety are raised as discussed in Chapter D5.12 by Moseley and Davies. Here special considerations apply since the laser is being used to produce controlled injury, rather than in the normal realm of laser safety in ensuring the absence of injury.

The application of lasers in medicine is progressing strongly in clinical procedures and in diagnostics. The steady development of new, flexible, smaller and more efficient laser sources, optical materials and techniques will bring forward further advanced medical systems. This can be anticipated to include new wavelengths useful for medicine, further use of non-linear optics to give frequency conversion and tunable sources, for example, based on OPOs operating over a broad wavelength range, and compact ultrashort pulse systems to give specific tissue interaction. The application of lasers in medicine and biology can be expected to continue to advance to realize its potential.

The strong interdisciplinary nature of the field of lasers in medicine and biology is reflected in the broad range of journals in which research is published. There are several journals which are devoted specifically to the subject and include:

Lasers in Surgery and Medicine, Lasers in Medical Science, Journal of Biomedical Optics, Lasers in Life Sciences, Medical Laser Applications, Journal of Clinical Laser Medicine and Surgery and *Journal of Oral Laser Applications.*

In addition, a number of journals publish substantially on relevant biology and medical topics. These include:

Photochemistry and Photobiology, Journal of Photochemistry and Photobiology, Applied Optics, IEEE Journal of Quantum Electronics, IEEE Journal of Selected Topics in Quantum Electronics (July/August 1999 and December 2001), *IEEE Transactions on Biomedical Engineering, Photochemical and Photobiological Sciences* and *Physics in Medicine and Biology.*

Aspects of the laser biomedical industry and current news are described in the magazines:

Biophotonics International, Laser Focus World and Photonics Spectra.

There are many national societies devoted to lasers in medicine. The American Society for Laser Surgery and Medicine (ASLMS) holds an annual meeting and also specialist meetings. In Europe there is the European Laser Association.

In addition to the many specialist medical and scientific conferences, the International Society for Optical Engineering, SPIE, holds annual Biomedical Optics (BIOS) Conferences in the United States and Europe. The published SPIE proceedings from these conferences provide an extensive and up-to-date literature.

14

Light–Tissue Interactions

Steven Jacques and Michael Patterson

CONTENTS

14.1 Introduction 201
14.2 Fundamental Interactions 202
  14.2.1 Absorption 202
  14.2.2 Luminescence 203
  14.2.3 Elastic Scattering 203
  14.2.4 Scattering by a Dielectric Particle 204
  14.2.5 Scattering by Many Particles 205
  14.2.6 Inelastic Scattering 205
  14.2.7 Miscellaneous 206
14.3 Optical Properties of Tissue and Their Measurement 207
  14.3.1 Absorption Coefficient 207
  14.3.2 Scattering Coefficient 207
  14.3.3 Scattering Phase Function 207
  14.3.4 Measurement of Optical Properties Ex Vivo 207
  14.3.5 Measurement of Optical Properties In Vivo 208
14.4 Physical Models of Light Propagation in Tissue 210
  14.4.1 The Radiative Transport Equation 210
  14.4.2 Numerical Solution of the RTE 210
  14.4.3 Monte Carlo Simulation 210
  14.4.4 The Diffusion Approximation 211
  14.4.5 Fluence Rate Distributions in Tissue 212
14.5 Therapeutic Laser–Tissue Interactions 213
  14.5.1 Introduction 213
  14.5.2 Photochemical Effects 213
  14.5.3 Photothermal Effects 214
    14.5.3.1 Laser Heating 215
    14.5.3.2 Thermal Diffusion 215
    14.5.3.3 Thermal Damage 216
  14.5.4 Photomechanical Effects 219
    14.5.4.1 Thermoelastic Expansion 219
    14.5.4.2 Spallation and Cavitation 221
    14.5.4.3 Vaporization 222
    14.5.4.4 Plasma Formation 224
14.6 Summary 224
References 224

14.1 Introduction

The application of lasers and other optical technology to problems in biomedicine is a rapidly growing field. These applications can be classified as diagnostic or therapeutic. In a diagnostic application, the goal is to learn something about the physiology or pathology of the tissue through its interaction with light. For example, the fluorescence emitted by tissue after laser excitation is characteristic of its chemical composition which, in turn, can depend on the nature and stage of certain diseases such as cancer [1]. The fundamental optical processes that can be exploited for diagnostic information are absorption, scattering (both elastic and inelastic) and luminescence (both fluorescence and phosphorescence). Most often optical diagnostics rely on interaction with endogenous tissue constituents but it is also possible to administer an exogenous agent that interacts with light. For instance, a bolus injection of indocyanine green, an infrared absorbing dye, has been used to measure blood flow in specific organs [2]. The interpretation of a diagnostic optical measurement may be based on empirical correlation with other data or it may require a thorough understanding of the particular light–tissue interaction being used as well as the propagation of light within the tissue.

When lasers are used for therapeutic intervention, the goal is permanent modification of the tissue. This can range from simple cutting associated with surgery to the initiation of cytotoxic chemical reactions in photodynamic therapy (PDT). In all cases electromagnetic energy must be absorbed by the tissue or by an exogenous agent. The subsequent processes can be classified as photochemical, thermal or photomechanical. The dominant process is determined by the rate at which energy is absorbed per unit volume of tissue. For linear absorption this quantity is simply the product of the tissue absorption coefficient and the local light fluence rate. Both of these can vary by many orders of magnitude so that a wide range of therapeutic results is possible. Optimization of these procedures requires knowledge of tissue optical properties and light propagation.

The purpose of this chapter is to provide an overview of the physics underlying both diagnostic and therapeutic applications. First, the fundamental processes of absorption, scattering and luminescence are defined and discussed. Our goal is not to provide a detailed description but a qualitative explanation with an emphasis on the aspects relevant to laser–tissue interactions. The absorption and scattering coefficients of tissue are defined and their magnitude and wavelength dependence discussed along with techniques to measure them *ex vivo* and *in vivo*. We then consider the problem of light propagation in tissue and explain the radiative transport equation (RTE). Solutions to this equation are considered with emphasis on Monte Carlo simulations and the diffusion approximation—two methods most often applied in tissue optics. Finally we discuss the photochemical, thermal and photomechanical processes relevant to therapeutic interventions.

14.2 Fundamental Interactions

14.2.1 Absorption

Clinical laser applications span a range of wavelengths from 193 nm (ultraviolet excimer laser) to 10.6 μm (infrared carbon dioxide laser). The equivalent photon energy span is 6.4–0.12 eV. Absorption of a photon by an atom or molecule is possible when its energy matches the difference in energy between discrete quantum states of the atom or molecule. These absorption processes (and subsequent events) are often depicted in a Jablonski diagram [3]. Figure 14.1 shows a typical diagram for an organic molecule. The singlet ground state is denoted by S_0 and the first and second singlet excited electronic states by S_1 and S_2, respectively. Vibrational levels 0,1,2, … are associated with each of the electronic states. In the diagram, electronic transitions associated with photon absorption are indicated by vertical arrows. For this hypothetical molecule, the first excited singlet state can be populated by absorption at four different wavelengths corresponding to transitions from the ground state to each of the vibrational levels. These absorption events occur on a time scale of about 10^{-15} s. Excitation to the higher vibrational levels is followed by a non-radiative decay to the lowest level.

The probability of absorption at a specific wavelength can be expressed as a molecular cross section, σ. Suppose the local photon fluence rate is H photons cm^{-2}s^{-1} and the molecular concentration is N cm^{-3}. The number of photons absorbed per unit volume per unit time is $H\,N\sigma$. The product $N\sigma$ has units cm^{-1} and is known as the linear absorption coefficient, μ_a. The reciprocal of the absorption coefficient is the mean free path for absorption, so that the probability of photon absorption over an infinitesimal path length dx is μ_a dx. A plot of the absorption coefficient as a function of wavelength is the absorption spectrum. (Note that spectroscopists prefer to use wavenumber as the abscissa because it is equivalent to energy but we will follow common practice in biomedical optics.) The absorption spectrum corresponding to the Jablonski diagram

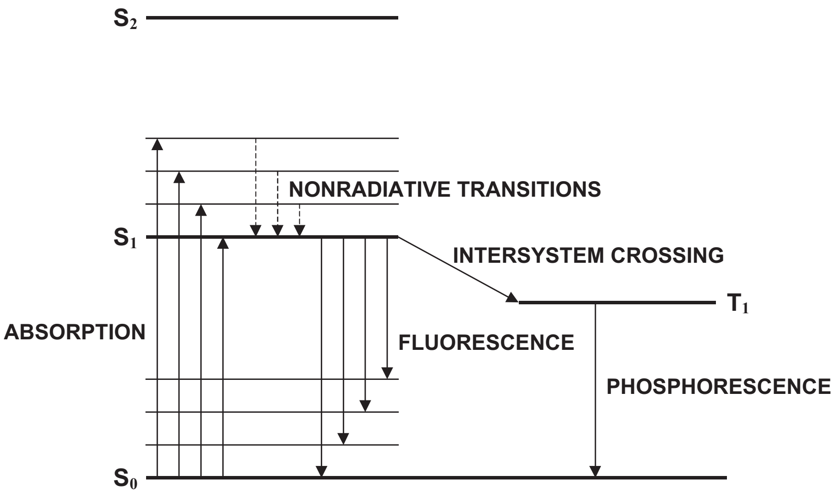


FIGURE 14.1 Typical Jablonski diagram for an aromatic organic compound. The processes of light absorption, fluorescence, phosphorescence and intersystem crossing are illustrated.

of Figure 14.1 is illustrated in Figure 14.2. The four transitions to the first excited singlet state result in four peaks in the absorption spectrum. These typically appear as broad peaks rather than sharp lines because of thermal motion, additional rotational energy levels and environmental effects. The relative amplitude of the peaks depends on detailed molecular structure and is beyond the scope of this chapter (see Ref. [4]). With sufficient photon energy, it also may be possible to excite the second singlet state and associated vibrational levels but that wavelength range has not been included in Figure 14.2. The spectrum in Figure 14.2 is typical of aromatic organic molecules in the ultraviolet, visible and near infrared. Beyond about 900 nm, excitation of vibrational modes is the primary absorption of mechanism. Specific examples of light-absorbing molecules in tissue and the magnitude of tissue absorption coefficients are discussed in Section 14.3.

At very high light fluence rates, multi-photon absorption is also possible. For example, if the first singlet state corresponds to a photon energy of 2 eV (corresponding wavelength 620 nm), the excitation can be achieved by simultaneous absorption of two photons, each of energy 1 eV (corresponding wavelength 1240 nm). Sufficiently high fluence rates can be achieved with pulsed lasers and focused excitation but light scattering limits multi-photon absorption to thin tissue slices or superficial layers. The principal application has been to multi-photon excitation for fluorescence microscopy. Because the fluorescence excitation is limited to the high fluence region, superior spatial resolution can be achieved.

14.2.2 Luminescence

The emission of light from excited electronic states is called luminescence. This includes the processes of fluorescence and phosphorescence. Fluorescence refers to emission associated with a spin-allowed electronic transition. For example, in Figure 14.1, the outer electron pair in the state S_1 has the *same* overall spin (zero) as that of the electrons in the ground state S_0 so the radiative transition indicated by the vertical line in Figure 14.1 is allowed by the spin conservation rule. Fluorescence emission occurs on a time scale of about 10^{-9} s. Transitions can also occur to the vibrational levels associated with S_0, so several peaks may be evident in the fluorescence emission spectrum. Because the spacing of vibrational energy levels associated with S_1 and S_0 is similar, the emission spectrum often appears as a 'mirror image' of the absorption spectrum, as illustrated in Figure 14.2. Because transitions occur from S_1, the shape of the emission spectrum is independent of the wavelength used to excite the molecule. However, the observed intensity of fluorescence can be a strong function of the excitation wavelength because of the wavelength dependence of the absorption. The probability of fluorescence photon emission after photon absorption is called the fluorescence quantum yield. Because the excited molecule can undergo non-radiative (or 'quenching') transitions to the ground state (e.g. by collision), the quantum yield can be highly dependent on the environment of the molecule.

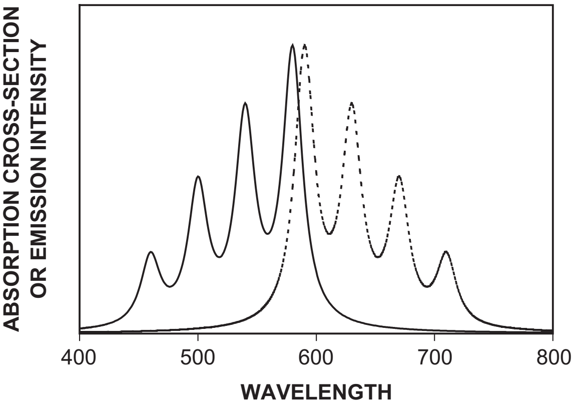


FIGURE 14.2 Absorption (full curve) and fluorescence emission (dotted curve) spectra corresponding to the Jablonski diagram in Figure 14.1. Note that the lowest energy absorption and highest energy emission do not occur at exactly the same wavelength due to interactions of the excited state with surrounding molecules.

Phosphorescence refers to the emission associated with forbidden transitions from excited states to the ground state where the total spin of the electron pair in the excited state T_1 (1 in this case) differs from that in the ground state S_0. These excited states are referred to as triplet states and can result from a rearrangement of electron spin in the S_1 state—a process referred to as inter-system crossing. The Jablonski diagram in Figure 14.1 shows the excited triplet state T_1 that results from inter-system crossing. Because the T_1–S_0 transition is forbidden, the triplet state lifetime can be very long, especially in the gas phase. In condensed media, the triplet is usually de-excited by energy transfer to surrounding molecules so that the effective triplet lifetime is of the order of 10^{-6} s. It is common for the triplet to transfer energy to molecular oxygen (a molecule with an unusual triplet ground state), so that the triplet lifetime of phosphorescent molecules has been used for *in vivo* measurement of oxygen concentration in tissue [5].

14.2.3 Elastic Scattering

Electromagnetic radiation exerts forces on the electrons in tissue thereby accelerating them. This acceleration results in radiation at the same frequency but in all directions. This is the origin of light scattering. In a completely uniform medium the secondary radiation interferes in such a way that only the incident electromagnetic wave is observable. In a medium where the index of refraction varies with position, however, scattered radiation can be detected. A casual glance through a microscope or at images of histological slides is enough to demonstrate that tissue is a medium in which the index of refraction changes over many different length scales ranging from cellular organelles much smaller than typical wavelengths, to macroscopic multi-cellular organization on a millimetre scale. Not surprisingly, light scattering by tissue is a complex phenomenon that is still an area of active research. To introduce important concepts of light scattering, we will first consider the much simpler case of an isolated dielectric particle. We will then introduce the complications of multiple particles, multiple scattering and dependent scattering.

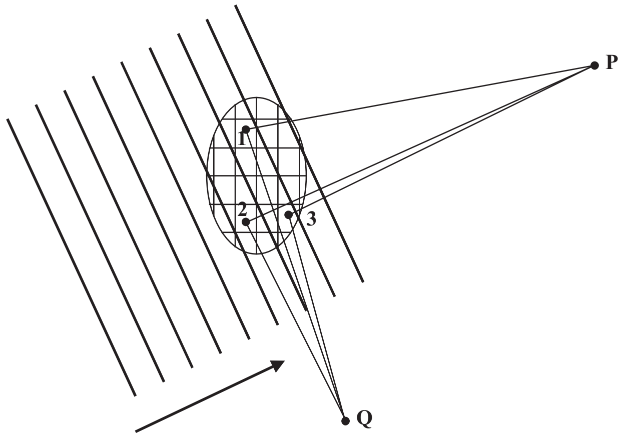


FIGURE 14.3 Scattering of an incident plane wave by a dielectric particle. The scattered radiation from each element is summed coherently at field points P and Q. A peak is observed in the forward direction (P) because scatter from elements such as 2 and 3 will be in phase, as will the contributions from 1 and 3.

14.2.4 Scattering by a Dielectric Particle

The scattering of a planar electromagnetic wave by a uniform sphere immersed in a medium with a different index of refraction is a problem solved by Gustav Mie in the early 1900s. We use it to define important parameters and to illustrate some principles. Rather than developing the exact mathematical solution of the problem, we follow the discussion in [6] to predict some qualitative results. Consider the case of an isolated particle irradiated by a plane wave illustrated in Figure 14.3. We can divide the particle into small regions in which a dipole moment is induced by the incident electromagnetic wave. The dipoles oscillate at the frequency of the incident wave and radiate in all directions. The scattered field at some arbitrary point, P, in Figure 14.3 is the superposition of the scattered waves from all of the regions of the particle and must incorporate the relative phase of each contribution. If the scattering particle is small relative to the optical wavelength, the distance from all regions of the particle to P is practically the same, so the scattered waves will be in phase. In this case we would anticipate that the scattering would be comparable in all directions. For particles with dimensions comparable to or larger than the wavelength, path length differences will cause substantial differences in the phase of the waves scattered from different regions. The interference of these waves could lead to a complex scattering pattern with maxima and minima in specific directions. Note that a maximum in the forward direction is always expected because displacements of scattering regions in the direction of the incident wave do not contribute a phase shift—a region that is farther from the source is correspondingly closer to P.

Before comparing features predicted on the basis of simple physical arguments with the exact Mie solutions, we define some additional quantities. The total energy scattered in all directions by the particle is equal to the energy of the incident wave falling on an area C_{sca}. This area is the called the scattering cross section of the particle. The ratio of the scattering cross section to the actual geometric cross section of the particle (projected onto a plane perpendicular to the direction of propagation of the incident wave) is called the scattering efficiency, Q_{sca}. We can also define a differential scattering cross section $\mathrm{d}C_{\text{sca}}/\mathrm{d}\boldsymbol{\Omega}$ which describes the amount of light scattered in a specific direction per unit solid angle. If the differential scattering cross section is normalized so that its integral over all directions is unity, the resulting function is called the phase function, $p(\boldsymbol{\Omega}', \boldsymbol{\Omega})$, that is,

$$\int_{4\pi} p(\boldsymbol{\Omega}', \boldsymbol{\Omega})\ d\Omega = 1. \tag{14.1}$$

The average cosine of the scattering angle is the anisotropy parameter, g:

$$g = \int_{4\pi} p(\boldsymbol{\Omega}', \boldsymbol{\Omega})(\boldsymbol{\Omega}' \cdot \boldsymbol{\Omega})\ \mathrm{d}\boldsymbol{\Omega}. \tag{14.2}$$

Note that isotropic scattering (i.e. the phase function is independent of direction) corresponds to $g = 0$, whereas the anisotropy parameter approaches 1 as the phase function becomes more peaked in the forward direction.

The exact Mie solutions for the differential scattering cross section depend on two variables: the size parameter, X, and the relative index of refraction, m, where

$$X = \frac{\pi D}{\lambda} \quad m = \frac{n_{\mathrm{p}}}{n} \tag{14.3}$$

where λ is the optical wavelength, D is the particle diameter, n is the index of refraction of the medium and n_{p} is the index of refraction of the particle. A particle that absorbs light will have a complex index of refraction but we will not consider that case here. Figure 14.4 shows the phase function for a sphere with $X \ll 1$. As expected, there is only a weak dependence of the scattering on azimuthal angle. Figure 14.4 also introduces an additional complication. In addition to the phase function for unpolarized light, the scattering diagrams for light linearly

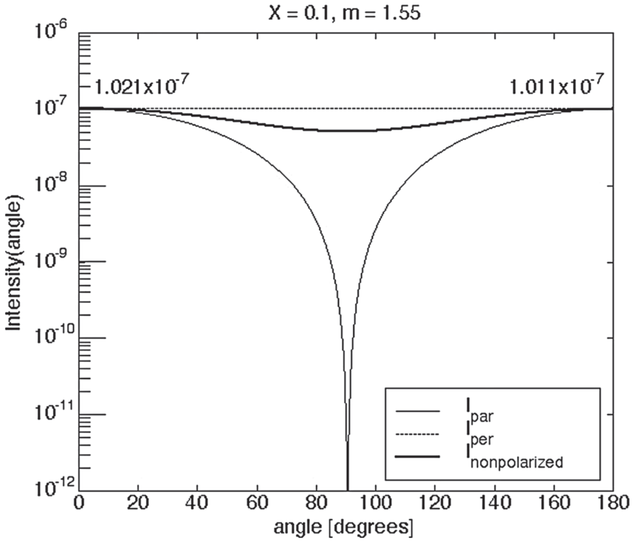


FIGURE 14.4 Angular dependence of light scattered by a non-absorbing dielectric sphere with $X = 0.1$ and $m = 1.55$. Results are shown for unpolarized incident light as well as light polarized parallel (par) and perpendicular (per) to the scattering plane. The units are steradian^{-1}.

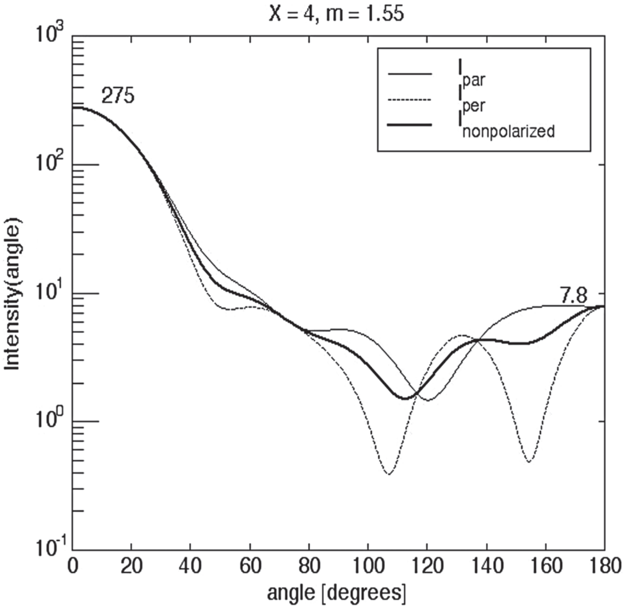


FIGURE 14.5 Angular dependence of light scattered by a non-absorbing dielectric sphere with $X = 4$ and $m = 1.55$. Results are shown for unpolarized incident light as well a light polarized parallel (par) and perpendicular (per) to the scattering plane. The units are steradian^{-1}.

polarized in the scattering plane and perpendicular to the scattering plane are also shown. Clearly, the polarization state of the scattered light can be different from that of the incident light—for example, even if the incident light is unpolarized, the light scattered at 90° is completely polarized perpendicular to the scattering plane. In general, the polarization state of light is completely described by the four Stokes parameters related to the intensity, the amount and direction of a linearly polarized component and the amount and helicity of a circularly polarized component. An optical interaction can change all four Stokes parameters, so a complete description of the interaction requires the 16 elements of the so-called Mueller matrix [6]. Figure 14.5 depicts the scattering phase function for a larger particle. The complex nature of the phase function due to interference effects is evident along with the expected forward peak.

14.2.5 Scattering by Many Particles

If the scattering particles are sparsely distributed so that each particle 'sees' only the incident wave, then the total scattered field is just the coherent sum of those from each particle. If the particles are stationary and the light source is coherent over a sufficiently long time, then this scattered field will reflect the interference of contributions from each particle and will display variations in intensity known as speckle. The spatial structure of the speckle pattern depends in a complex way on the spatial arrangement of the particles. A familiar example of a speckle pattern can be observed in the diffuse reflection of a laser beam from a sheet of white paper. If the scattering particles are moving, then the relative phase of the waves scattered from individual particles will change with time and so will the resulting speckle pattern. The characteristic time over which the pattern changes are related to the average speed of the particles. If the measurement time is long compared to this characteristic time, a speckle pattern will not be observed. For measurement times short compared to the characteristic time, large variations in intensity will be observed if the detector is comparable in size to the speckle dimensions. As the size of the detector increases, the variation is reduced because an increasing number of speckles contributes to the signal.

If there are N identical scattering particles per unit volume and each has a scattering cross section C_{sca}, we can define a linear scattering coefficient $\mu_s = NC_{sca}$. The reciprocal of the scattering coefficient is the mean free path between scattering events and the probability of photon scatter over an infinitesimal path length dx is $\mu_s\, dx$. Note that the linear scattering coefficient is a property of the medium (i.e. particles and the material in which they are suspended) as a whole. In this simple case, there is a straightforward relation between μ_s and the properties of individual particles. This need not be the case—the scattering coefficient can be defined and measured for much more complex media. For example, if the particles are densely packed (spacing less than a few particle diameters), they will not scatter independently and μ_s will not be simply the product of N and C_{sca}.

The fluence rate due to unscattered light from a beam incident on a scattering medium, such as a suspension of nonabsorbing particles, will be reduced with distance x into the medium according to

$$\psi(x) = \psi_0 \exp(-\mu_s x). \qquad (14.4)$$

At sufficient depth, the incident beam will be attenuated to such an extent that the major contribution to the fluence rate will be from scattered radiation. Under these conditions, radiation which has been scattered more than once, so-called multiple scattered light, will be significant. As discussed later, this is often the case in tissue unless the absorption coefficient is very high.

14.2.6 Inelastic Scattering

In the preceding discussion, we assumed that the frequency of the scattered electromagnetic wave is identical to that of the incident wave. There are two important situations where this is not the case. If light is scattered by a moving particle, as shown in Figure 14.6, the scattered wave will be Doppler shifted by an amount

$$\Delta f = \frac{vf}{c}(\cos\theta + \cos\varphi) \qquad (14.5)$$

where f is the frequency of the incident light, v is the speed of the particle and θ and φ are the angles indicated in Figure 14.6. For a collection of moving scattering particles, a spectrum of Doppler-shifted light would be observed which can be related to the distribution of particle velocities. Intuitively, it can be seen that there must be a relationship between this spectrum and the temporal variations in the speckle discussed earlier. This relationship is well understood but will not be explicitly developed here (see Ref. [7]). The obvious biomedical

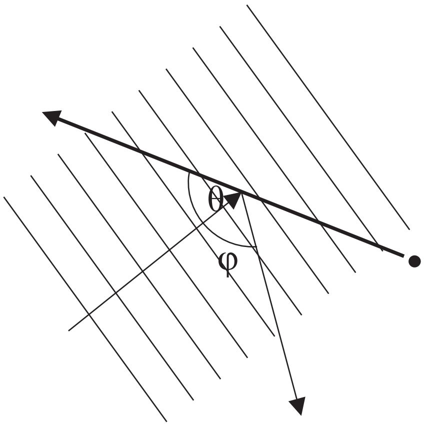


FIGURE 14.6 Geometry for calculation of the Doppler shift in equation (14.5). For simplicity it is assumed that the particle motion is confined to the scattering plane.

application of this ‘quasi-elastic light scattering’ is to the measurement of blood flow in tissue.

Inelastic scattering also occurs when the electromagnetic wave excites molecular vibrations. If a molecule has a vibrational state at energy E above the ground state, a photon of energy hf can be inelastically scattered at energy $hf - E$. This phenomenon is referred to as Stokes Raman scattering. It is also possible for the photon to interact with a molecule already in the higher vibrational state so that it gains energy E. This is referred to as anti-Stokes Raman scattering. Irradiation of a sample with a single-frequency laser will lead to a Raman spectrum of inelastically scattered light. The nature of this spectrum will depend on the abundance of specific molecules in the samples and Raman spectroscopy has evolved as a powerful tool for analytical chemistry. The probability of Raman scattering is usually orders of magnitude less than that of elastic scattering so that *in vivo* applications have only begun to emerge with technological improvements in detectors, lasers and optical fibre delivery systems [8]. Figure 14.7 shows an example of a Raman spectrum obtained *in vivo* from the GI tract.

14.2.7 Miscellaneous

For completeness we list here some other interactions which can change the characteristics of a light wave as it propagates through tissue and that may allow characterization of the tissue under some circumstances. Different tissues have different indices of refraction and these can be measured directly by looking at the speed of light propagation. Some tissues display birefringence and this property has been characterized by examination of the polarization state of light transmitted through the tissue [9]. Finally, tissue may contain optically active molecules such as glucose. Attempts have been made to measure glucose concentration in the anterior chamber of the

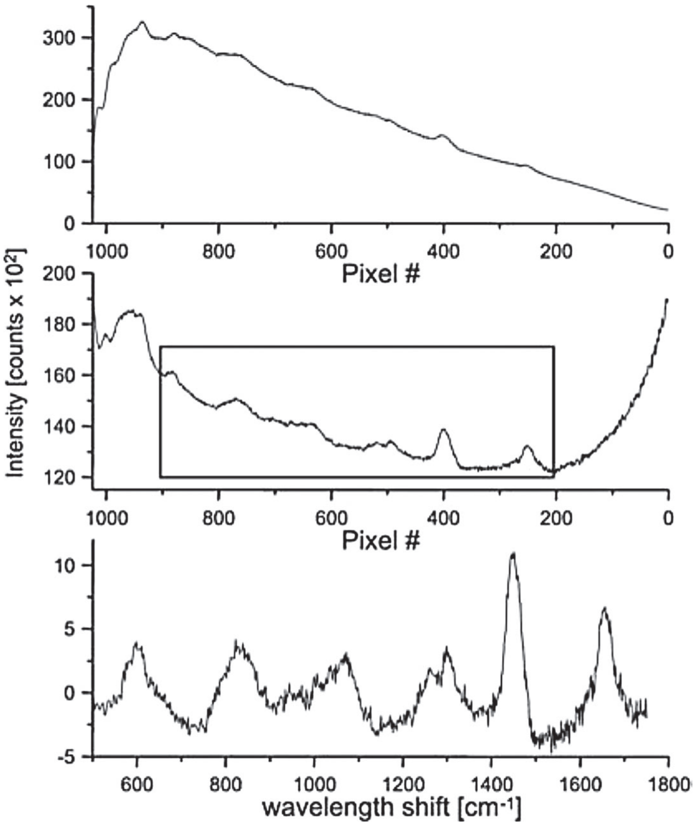


FIGURE 14.7 Raman spectrum obtained *in vivo* from a colon polyp using a fibre optic probe via an endoscope. The top panel shows the raw data which have been corrected for the system response in the middle panel. In the bottom panel wavelength calibration and background subtraction have been applied to the boxed region. The excitation wavelength was 785 nm. (Spectrum courtesy of Martin Shim, Hamilton Regional Cancer Centre.)

eye by measurement of the polarization state of light that has traversed the chamber [10].

14.3 Optical Properties of Tissue and Their Measurement

In the previous section, we defined the absorption coefficient, scattering coefficient and scattering phase function. In this section we discuss the magnitude of these quantities for soft tissues and their variation with wavelength. These optical properties can be of intrinsic interest in diagnostic procedures (e.g. absorption spectroscopy) and they are the determinants of the fluence rate distribution in tissue during therapeutic interventions. We will not discuss properties such as birefringence because they are not yet as well characterized, do not have a significant effect on fluence rate distributions and can only be measured in superficial tissues *in vivo*. We complete this section with a brief description of methods for the measurement of tissue optical properties *ex vivo* and *in vivo*.

14.3.1 Absorption Coefficient

The linear absorption coefficient of tissue can be determined from the concentration and molecular cross section of all chemical species present. These, in turn, will have specific spectra so that a chromophore that dominates absorption at one wavelength may make an insignificant contribution at other wavelengths. In the ultraviolet region of the spectrum, tissue absorption is dominated by proteins. Not only is the concentration of proteins high, the amino acids phenylalanine, tyrosine and tryptophan are efficient absorbers in the 250–300 nm range. At even shorter wavelengths, the peptide bond itself makes a large contribution. As the wavelength increases, protein absorption decreases rapidly and other chromophores, especially haemoglobin, become dominant. The individual absorption peaks of haemoglobin and the change in spectrum with oxygenation can be readily observed *in vivo*. The infrared absorption of tissue is dominated by its most abundant chromophore, water. The vibrational modes of the O–H bond produce a series of peaks with the maximum occurring at 2.95 μm. Figure 14.8 shows the absorption coefficient of the three principal tissue chromophores at 'typical' concentrations. Of course their concentration will vary from tissue to tissue and other pigments may be significant in organs such as the liver and the skin but Figure 14.8 correctly predicts the order of magnitude of tissue absorption and the main features of its dependence on wavelength.

14.3.2 Scattering Coefficient

A simple approach analogous to that taken earlier to predict the nature of tissue absorption is not obvious for scattering. Tissue is a complex medium in which the index of refraction varies in a complex way over different length scales in different tissues. Some progress has been made in developing models of tissue scattering that incorporate fractal distributions of scattering particles of different size ranges and the predictions of these models are in reasonable agreement with measured data [12]. We will not attempt to develop these arguments here but will state some accepted features of tissue scattering. First, the scattering coefficient is very high when compared to, say, low-energy X-rays. Typical values at visible wavelengths are 10–50 mm^{-1}, corresponding to a mean free path of 20–100 μm. In general, the scattering coefficient is observed to decrease monotonically as the wavelength increases although few studies have been performed over a wide spectral range. Except in the ultraviolet and infrared, the scattering coefficient is orders of magnitude larger than the absorption coefficient. This relationship is depicted in Figure 14.8 and is a major determinant of the fluence rate distribution that results in tissue upon laser irradiation.

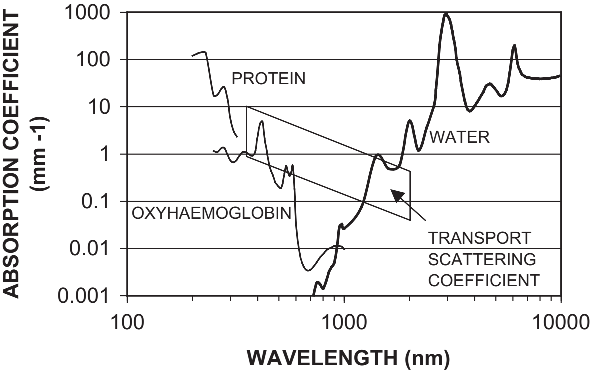


FIGURE 14.8 Absorption spectra of three important tissue constituents: protein (bovine serum albumin at 170 mg g^{-1}), oxyhaemoglobin (40 μM) and water (40 M). The parallelogram encompasses reported values for the transport scattering coefficient of soft tissues [11].

14.3.3 Scattering Phase Function

Even though small structures contribute to tissue scattering, those comparable to or larger than the optical wavelength dominate the phase function. As previously discussed, this leads to scattering that is forward peaked, typical values of the anisotropy parameter for tissue are 0.9 or higher in the visible. Typically no other peaks are observed in the phase function (presumably because of the range of scatterer sizes and their irregular shapes) although some experiments show a small backscattering enhancement. Figure 14.9 shows the measured phase function for porcine brain tissue at 633 nm.

14.3.4 Measurement of Optical Properties Ex Vivo

We can classify the measurement of fundamental tissue optical properties (i.e. μ_a, μ_s, phase function) as direct or indirect. Direct methods yield the quantity of interest without resort to a model of light propagation in the tissue sample. Indirect methods utilize measurements of quantities (such as diffuse reflectance) that depend on the fundamental properties, often in a complex way. A forward model of light propagation (see Section 14.4) in the tissue sample can be used to calculate the dependent quantity for a given set of fundamental optical properties. Estimation of the fundamental properties is an inverse problem solved by the following strategy: first the optical

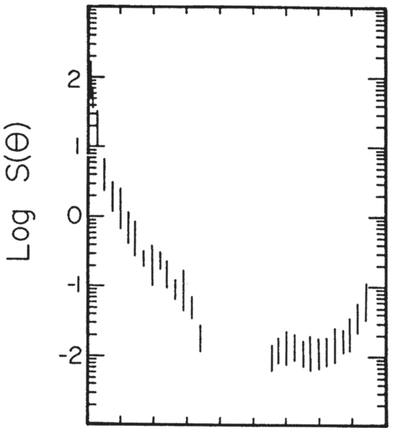


FIGURE 14.9 Measured scattering phase function for porcine brain tissue at 633 nm [13]. Results were obtained for three 20 µm thick samples. Vertical bars represent the range of values for the three samples. The average cosine of the scattering angle is 0.94.

properties of the sample are estimated using whatever *a priori* information is available. The forward model is then used to calculate the dependent quantities for that set of properties. The calculations are compared with actual measurements and the estimates are altered appropriately. The entire process is repeated until measurements and calculations agree to some tolerance yielding the best estimate of the fundamental optical properties.

Direct measurement of the scattering phase function is possible for a tissue sample that is less than one mean free path thick. In thicker specimens multiple scattering and/or absorption distort the measured scattering pattern. The measurements in Figure 14.9 were obtained on tissue slices approximately 20 µm thick sandwiched between glass slides. A laser beam was incident on the sample and a collimated detector was rotated about the sample to measure the phase function [13]. In these measurements, it is necessary to eliminate or correct for refraction at the interfaces. It is also possible to perform direct measurements of the total attenuation coefficient, $\mu_t = \mu_s + \mu_a$. As shown in Figure 14.10, a narrow collimated beam is incident on the sample and a detector measures unscattered and unabsorbed light after transmission. It is not necessary for the sample to be optically 'thin' but steps must be taken to ensure that the detector does not collect a significant amount of scattered light. This can be achieved by geometric collimation, time-gating for an incident pulse, coherence gating or a combination of these techniques. The attenuation coefficient is typically at least 10 mm^{-1}, so that a practical limit on sample thickness is imposed by the signal/noise or signal/scattered background ratio. Strictly speaking μ_t is not a fundamental optical property as it does not discriminate between absorption and scattering. However, as indicated in Figure 14.8, there are spectral regions where $\mu_s \gg \mu_a$ or $\mu_a \gg \mu_s$ so that one or the other may be measured directly. It is also possible to use coherence gating to detect light that is singly scattered in the backward direction and to measure the depth at which that backscatter occurs. This is the essence of optical coherence tomography (OCT [14]), an imaging technique analogous to B-mode ultrasound imaging. The technique is mentioned here because the decrease in the backscatter signal with depth is a direct measure of the total attenuation coefficient and the measurement can be performed on thick samples.

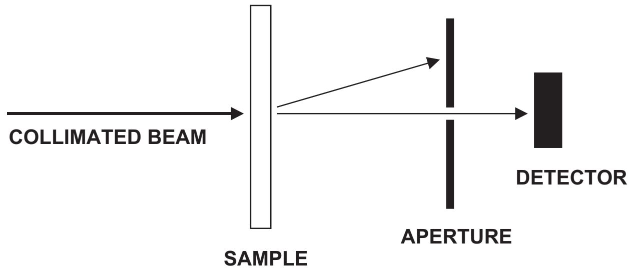


FIGURE 14.10 Apparatus for measurement of the total attenuation coefficient, μ_t.

A large variety of indirect methods is potentially applicable to tissue samples as the spatial and temporal characteristics of both transmitted and reflected light can be utilized. The strategy most often applied is to measure the diffuse reflectance and transmittance of a tissue slab as shown in Figure 14.11 [15]. The two integrating spheres have a high diffusely reflecting coating so that the signal measured by the two detectors is proportional to the total diffuse reflectance, *R*, or transmittance, *T*, of the sample. Mathematical models are used to infer the fundamental optical properties from *R* and *T*. A common problem with indirect methods is the failure to separate the phase function and the scattering coefficient. A sample with low μ_s and low *g* will result in the same *R* and *T* as a sample with high μ_s and high *g*—an example of similarity. Mathematically, similar results are expected when the reduced (or transport) scattering coefficient, $\mu_s' = (1-g)\mu_s$, is identical for two samples. If it is important to distinguish μ_s and *g*, it may be necessary to incorporate a direct measurement on the sample as described earlier.

14.3.5 Measurement of Optical Properties In Vivo

The preparation of tissue samples introduces the possibility of artefacts caused by desiccation, blood loss and freezing and thawing if these are used in the process. For these reasons, measurements *in vivo* are preferable but they present their own challenges. The direct methods described earlier can rarely

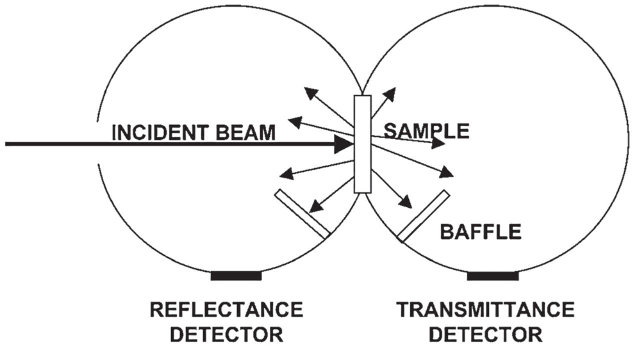


FIGURE 14.11 Apparatus for the simultaneous measurement of diffuse reflectance and transmittance from a tissue sample *ex vivo*. The baffles prevent the detection of light that has not been diffusely reflected by the integrating sphere.

be used because unscattered transmitted light is not detectable for tissues thicker than a few millimetres. OCT methods can be used to measure μ_t in a backscatter geometry but only for superficial tissues such as the skin. Therefore, most *in vivo* measurements are indirect and most are performed in reflectance geometry, although transmission measurements are possible in some sites such as the breast. Interstitial measurements, where light sources and detectors are inserted into the tissue are also possible but these introduce complications of uncertainty in location and artefacts due to local bleeding. The most successful approach to date is illustrated in Figure 14.12. An optical fibre is used to deliver light to the tissue surface and additional fibres detect diffusely reflected light at one or more well-defined locations on the tissue surface. If the tissue is uniform, the spatial dependence of the reflectance may be used to deduce the absorption and reduced scattering coefficients. However, tissue often has a layered structure and this method will not work well because light detected near the source has not penetrated as deeply as light detected farther away (see Figure 14.12). The deeper tissue may be characterized by measuring the distribution of times-of-flight from the source to the detector rather than just the amplitude of the signal. This can be done by using a picosecond pulsed laser as the source and time-correlated single-photon detection. Equivalently, the intensity of the source can be modulated at radio frequency (typically 100 MHz) and the phase and modulation of the detected light recorded. Figure 14.13 illustrates the broadening of a 40 ps pulse (751 nm) after propagation through 2 cm of muscle tissue in the forearm. Both μ_s' and μ_a of the muscle may be determined from the shape of this pulse. The analysis is relatively simple if the tissue optical properties are uniform. If this is not the case, it may be possible to estimate μ_a and μ_s' as functions of position in the tissue. This is the goal of diffuse optical tomography (DOT), a technique most highly developed for mapping the optical properties of the breast. Figure 14.14a illustrates the apparatus required to obtain phase and amplitude measurements at 16 different locations on the surface of the breast for 16 different locations of an intensity modulated source. Figure 14.14b shows a map of the absorption coefficient throughout a slice of the breast estimated using a diffusion model of light propagation in the breast. Further improvements in optical property measurement *in vivo* may be possible by combining optical measurements with anatomical information obtained by other imaging modalities such as magnetic resonance.

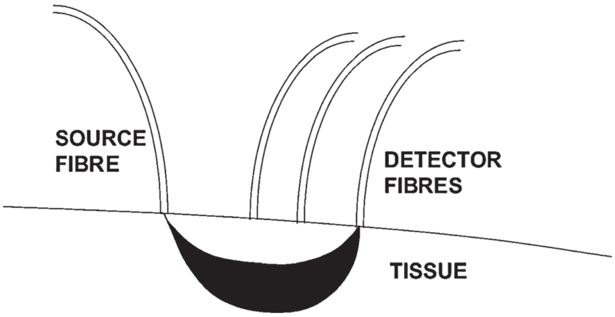


FIGURE 14.12 Typical set-up for the measurement of tissue optical properties *in vivo*. Light is delivered to the surface by an optical fibre and additional fibres detect scattered light at different positions. The shaded regions indicate schematically the distribution of photon paths from source to detector. Light detected close to the source has not penetrated as deeply as that detected at greater distances.

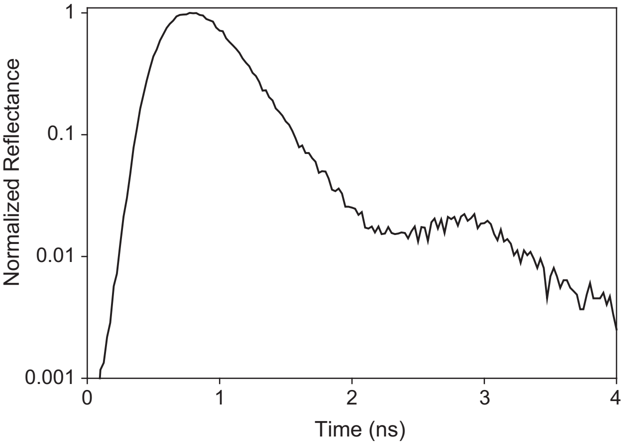
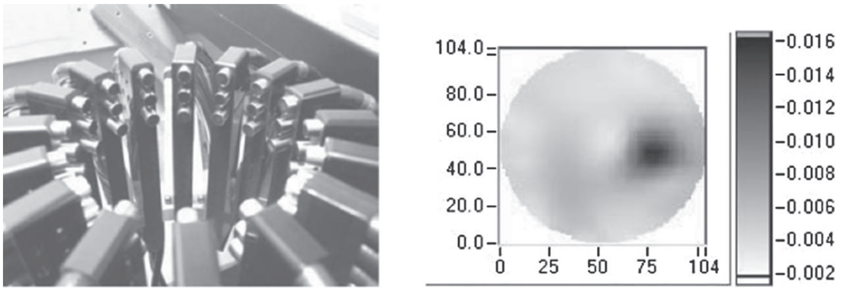


FIGURE 14.13 Broadening of 40 ps 751 nm pulse after transmission through 20 mm of muscle tissue *in vivo*.



(*a*) (*b*)

FIGURE 14.14 (a) Apparatus for DOT of the breast. In each of three planes, a total of 240 measurements are made of the phase and amplitude of modulated light transmitted through different portions of the breast. The modulation frequency is 100 MHz and four discrete wavelengths are used. Total data acquisition time is 30 s per wavelength. (b) Reconstructed image of the absorption coefficient at 761 nm for a subject with a 3.5 cm invasive ductal carcinoma of the breast. The units of the grey scale are mm^{-1} and the distance scale is in mm. (Photograph and image courtesy of Brian Pogue, Dartmouth College.)

14.4 Physical Models of Light Propagation in Tissue

As previously discussed, a physical model of light propagation in tissue may be required for several reasons: prediction of fluence distributions during therapeutic procedures, the interpretation of results from a diagnostic measurement, derivation of fundamental tissue optical properties or modelling for the optimization of both diagnostic and therapeutic applications. In this section we will discuss the most widely used model and aspects of its mathematical solution.

14.4.1 The Radiative Transport Equation

In most applications, the wave properties of the electromagnetic field are not important and we require only a model that predicts the transport of radiant energy through the tissue. By considering the energy balance in a small tissue volume it is possible to derive the RTE. In this equation the fundamental quantity is the radiance, $L(\boldsymbol{r}, \boldsymbol{\Omega})$, a function of position, $\boldsymbol{r}$, and direction, $\boldsymbol{\Omega}$. In general, the radiance is also a function of photon energy or wavelength but we will initially restrict our attention to monochromatic irradiation. The radiance is the energy transported per unit time per unit solid angle per unit area (perpendicular to the specific direction). Ignoring inelastic scattering and luminescence, the radiance can be decreased by absorption and elastic scattering (i.e. ‘out-scattering’) in the volume and increased by radiation sources and elastic scattering of radiation from other directions into the direction of interest (i.e. ‘in-scattering’). The detailed derivation can be found in texts such as [16] and the resulting time-dependent RTE is shown in equation (14.6)

$$\begin{aligned}\frac{1}{c}\frac{\partial}{\partial t}L(\mathrm{r},\boldsymbol{\Omega})+\boldsymbol{\Omega}\cdot\nabla L(\mathrm{r},\boldsymbol{\Omega})+[\mu_{\mathrm{a}}(\mathrm{r})+\mu_{\mathrm{s}}(\mathrm{r})]L(\mathrm{r},\ \boldsymbol{\Omega})&=\mu_{\mathrm{s}}(\mathrm{r})\\ \times\int_{4\pi}p(\boldsymbol{\Omega}',\boldsymbol{\Omega})L(\mathrm{r},\boldsymbol{\Omega}')\,\mathrm{d}\boldsymbol{\Omega}'+S(\mathrm{r},\boldsymbol{\Omega})&\qquad(14.6)\end{aligned}$$

where the linear absorption coefficient μ_{a}, the linear scattering coefficient μ_{s} and the phase function $p(\boldsymbol{\Omega}', \boldsymbol{\Omega})$ have been defined in Section 14.2. Note that they are written as spatially dependent quantities because, in the general case, propagation through different tissue types may be involved. The term $S(\boldsymbol{r}, \boldsymbol{\Omega})$ represents all sources of radiant energy within the tissue such as implanted optical fibres. It is often necessary to consider the behaviour of the radiance at boundaries between different types of tissue or between the tissue and other media such as air. This behaviour is governed by the Fresnel equations describing reflection and refraction at such boundaries. Mathematically, these lead to boundary conditions that must be imposed on the solution to equation (14.6). We note that it is also possible to use the RTE formalism to include polarization by writing a transport equation for each of the four Stokes components.

14.4.2 Numerical Solution of the RTE

Although equation (14.6) is relatively easy to understand and derive, it cannot be solved analytically for most cases of practical interest. In the fields of reactor physics and atmospheric physics, the RTE is often solved by numerical methods but these have been applied relatively infrequently in biomedical optics. Two general strategies may be employed. The first, referred to as the method of ‘discrete ordinates’ or the ‘multiflux method’, considers the radiance in specific solid angle ‘bins’. Equation (14.6) can then be replaced by a series of equations, each describing the radiance in a specific direction. The ‘in-scattering’ integral in equation (14.6) is replaced in each equation by a finite sum over the other directions. The second strategy is to expand the radiance as a sum of basis functions such as the spherical harmonics. Again this replaces equation (14.6) with a series of equations for each of the harmonics. The problem with either approach is that a large number of terms must be retained for accuracy if the radiance is highly anisotropic. This is often the case near boundaries, near anisotropic sources or in cases where absorption is high. In such instances it may be more efficient to use a Monte Carlo simulation.

14.4.3 Monte Carlo Simulation

The goal of Monte Carlo simulations is to estimate the radiance or other quantities by explicitly tracking the history of small ‘packets’ of radiant energy as they propagate through the tissue. In keeping with common practice, we will refer to these packets as ‘photons’ but the physics is consistent with the classical approach of the RTE. To illustrate the concepts, consider the simple case of Figure 14.15 where a pencil beam of radiation is incident on a uniform scattering medium and where the index of refraction of the scattering medium and its surroundings are identical. We consider an individual photon or energy packet after it enters the scattering medium. This photon will penetrate the tissue to some distance before its first

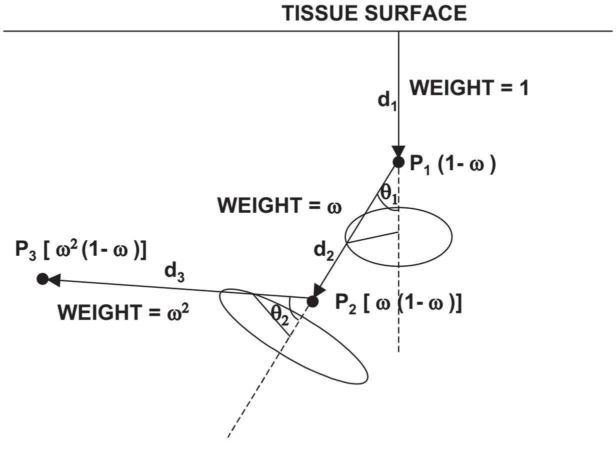


FIGURE 14.15 Schematic illustration of the first few steps in a Monte Carlo photon history. At each interaction site, the new direction of propagation is determined by sampling the polar and azimuthal scattering angles from appropriate probability distributions. The photon weight is reduced to account for energy deposited.

interaction. We know the probability distribution of this distance, *d,* is exponential according to

$$P(d) = \mu_t e^{-\mu_t d} \tag{14.7}$$

We can simulate this photon flight by sampling d from the probability distribution in equation (14.7). Efficient methods for sampling such distributions can be found in [17]. Once a value for d has been chosen, the location of the first photon interaction is determined, as shown in Figure 14.15. Ignoring low-probability events such as Raman scattering, this interaction is either absorption or elastic scattering. The choice can be made by random sampling from a uniform distribution on the interval (0,1). If the result is less than $\omega = \mu_s/(\mu_s+\mu_a)$, a ratio known as the single scattering albedo, then a scattering event takes place. Otherwise, the photon history is terminated and all of the photon energy is absorbed at this location. More commonly, a technique called survival weighting is used, where a fraction of the photon energy $(1-\omega)$ is deposited and the weight of the surviving photon is multiplied by ω. In order to continue the photon history it is necessary to determine the direction of propagation after scattering. This requires sampling the azimuthal angle from a uniform distribution (see Figure 14.15) and the polar angle, θ, from the appropriate phase function. For tissue, this could be a measured phase function such as that shown in Figure 14.9. Instead the Henyey–Greenstein phase function

$$p(\theta) = \frac{1-g^2}{(1+g^2-2g\cos\theta)^{3/2}} \tag{14.8}$$

is often used because it is mathematically simple and its single forward peak is a reasonable approximation to measured data.

Once the new direction of propagation is determined, as shown in Figure 14.15, the process of determining the next interaction site is repeated. This process is continued until the photon is scattered out of the tissue or the history is terminated according to other criteria. Typically, this termination occurs when the photon weight is low enough or the location so far from the region of interest that no significant contribution can be made to the scored quantity of interest. It may be necessary to perform test runs to determine criteria for termination. A safe way to end histories without introducing bias is a 'Russian roulette'. Once the photon weight falls below a pre-set value, a random number between zero and one is obtained. If the number is less than 0.5, the history is terminated. If the number is greater than 0.5, the photon weight is doubled and the history continues until the weight once again falls below the threshold. (Values other than 0.5 can be used with appropriate adjustment of the weight.)

Fresnel boundary conditions are readily incorporated into Monte Carlo simulations by splitting weights between reflected and transmitted photons. Regions with different optical properties can also be introduced, although the 'bookkeeping' required to track photons can be cumbersome. The chief disadvantage of Monte Carlo simulations is that the variance in any scored quantity depends on the number of photon histories that contribute. Even if millions of photon histories are traced, situations readily arise in which no photon histories contribute to the quantity of interest. For example, in Figure 14.12, for tissue optical properties typical of muscle in the near infrared, a source-detector distance of 20 mm and a 0.1 mm^2 detector, only one in 10^7 incident photons reaches the detector and yet these signals can be readily measured *in vivo.* To overcome this limitation, a number of 'variance reduction' techniques are often used in Monte Carlo simulations. The essence of these techniques is to bias the sampling of random variables to favour a selection that will increase the contribution of the photon history to the quantity of interest. After each biased sampling the photon weight is adjusted to 'correct' for the bias. The survival weighting and Russian roulette methods described earlier can be regarded as simple variance reduction techniques.

14.4.4 The Diffusion Approximation

In Figure 14.8 we demonstrated that there is a spectral region where scattering is the dominant light–tissue interaction. Under such conditions, and as long as boundaries and sources are not too close, the radiance is nearly isotropic. It can be expressed as the sum of an isotropic term and a direction dependent term as

$$L(\mathrm{r}, \boldsymbol{\Omega}) = \frac{1}{4\pi}\psi(\mathrm{r}) + \frac{3}{4\pi}\boldsymbol{\Omega}\cdot \mathrm{J}(\mathrm{r}). \tag{14.9}$$

If this expression is substituted in the RTE, it is possible to derive the diffusion equation for the fluence rate:

$$\frac{1}{c}\frac{\partial}{\partial t}\psi(\mathrm{r}, t) - \frac{1}{3\mu_s'}\nabla^2\psi(\mathrm{r}, t) + \mu_a\psi(\mathrm{r}, t) = s(\mathrm{r}, t). \tag{14.10}$$

This equation can be solved analytically for a number of cases of practical interest. A very useful result is the solution for an isotropic impulse source in an infinite uniform medium:

$$\psi_i(\mathrm{r}, t) = \left(c\frac{4\pi c}{3\mu_s'}t\right)^{-3/2} \exp\left(-\frac{3\mu_s' r^2}{4ct} - \mu_a ct\right). \tag{14.11}$$

Note that far from the source the fluence rate for a cw source decreases exponentially with distance according to an 'effective attenuation coefficient'

$$\mu_{\mathrm{eff}} = (3\mu_a\mu_s')^{1/2}. \tag{14.12}$$

This is a common feature of diffusion solutions to the RTE. The Green function in equation (14.11) can be used to construct the solution to other problems. For example, consider the problem in Figure 14.12 and suppose we seek a solution for the fluence rate in the tissue. The relatively narrow beam from the optical fibre source will be scattered as it enters the tissue. In fact, most of these interactions will be forward scattering but we can approximate by stating that all photons are isotropically scattered at a depth $1/\mu_s'$. This is the 'real' source of photons. We must also deal with the loss of photons at the tissue boundary that will reduce the fluence in the tissue. For this illustration, we will use the simplest possible boundary condition that the fluence rate vanishes at the boundary. A simple way to impose this condition is to add a negative 'image' source of photons a distance $1/\mu_s'$ outside the tissue boundary. The solution for the fluence rate in the tissue can now be expressed as the sum of contributions from the real and image sources as:

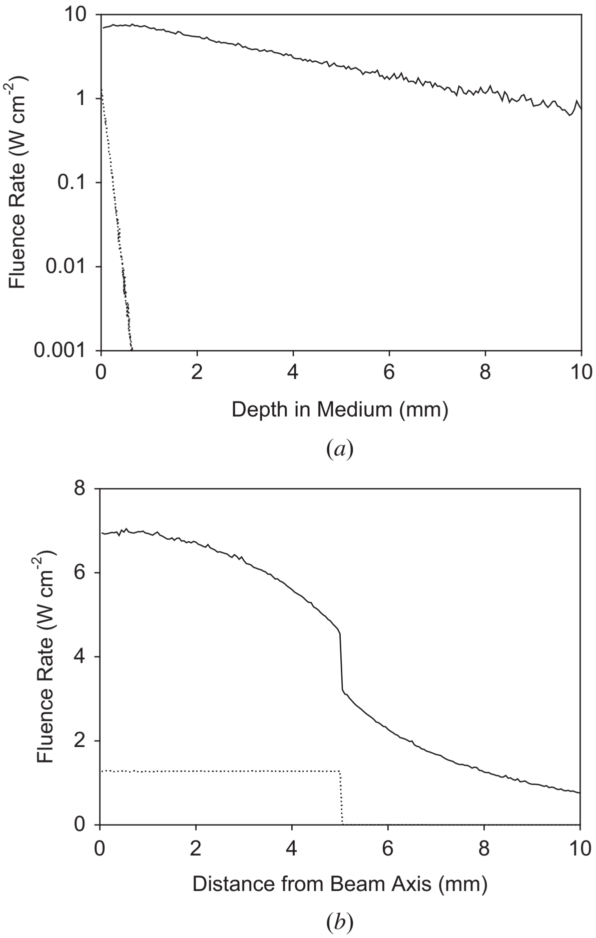
$$\psi(\mathrm{r},\ t)=\psi_{\mathrm{i}}\left(\left[\frac{1}{\mu_s'}-\mathrm{r}\right],\ t\right)-\psi_{\mathrm{i}}\left(\left[-\frac{1}{\mu_s'}-\mathrm{r}\right],\ t\right). \tag{14.13}$$

While similar ideas can be used in other cases, it may be necessary to solve equation (14.10) numerically. In biomedical optics, this has been most often done by using finite element methods. For example, the results in Figure 14.14b were obtained by employing a finite element model (FEM) with 2000 nodes as the forward solver of the diffusion equation. In general, the diffusion model approximation has proved to be extremely useful in advancing therapeutic and diagnostic laser applications.

14.4.5 Fluence Rate Distributions in Tissue

To summarize the preceding discussion on tissue optical properties and light propagation and as a preamble to the following section, it is instructive to examine the fluence rate distribution in tissue at different wavelengths. For a given irradiation geometry and irradiance, the resulting fluence rate distribution is governed by the scattering and absorption coefficients of the tissue. As discussed in Section 14.3 and illustrated in Figure 14.8, the absorption coefficient varies by several orders of magnitude across the spectrum of biomedical applications whilst the reduced scattering coefficient is a slowly decreasing function of wavelength. In the infrared and ultraviolet parts of the spectrum we would expect scattering to play a relatively minor role in determining the fluence distribution but in the visible and near infrared it should have a major influence. We will illustrate these differences by considering irradiation of a uniform tissue at 300 nm in the UV, where protein absorption is high, and at 700 nm, where tissue absorption is usually minimal. For the purposes of this illustration, let us assume that the reduced scattering coefficient is 1.0 mm^{-1} at both wavelengths, the absorption coefficient is 10 mm^{-1} at 300 nm and 0.005 mm^{-1} at 700 nm, the anisotropy parameter is 0.9 and the tissue is irradiated with a 10 mm diameter beam normally incident with a total power of 1 W (i.e. the irradiance is 1.27 W cm^{-2}). A Monte Carlo simulation was used to calculate the fluence rate distribution assuming an index of refraction of 1.4 for the tissue and 1.0 for the air above the tissue. Figure 14.16 plots the fluence rate on the central axis of the beam and across the beam just below the tissue surface.

The distributions are remarkably different. For 300 nm irradiation, tissue scattering has almost no influence. The edges of the beam are very sharp and the fluence rate falls off with depth according to exp − ($\mu_a \times$depth). The fluence rate just below the surface is equal to the irradiance. For irradiation at 700 nm, a number of different features are evident. In the 'diffusion regime', the fluence rate falls off approximately exponentially.



(*a*)

(*b*)

FIGURE 14.16 (a) Monte Carlo estimates of fluence rate as a function of depth on the central axis of a 1 cm diameter beam of irradiance 1.27 W cm^{-2} incident on 'tissue' at 300 nm (dotted line) and 700 nm. Note the larger uncertainty in the estimates at greater depths where fewer photon histories contribute to the scored fluence rate. (b) Monte Carlo estimates of fluence rate as a function of distance from the central axis just below the tissue surface. Results are shown for a 1 cm diameter beam of irradiance 1.27 W cm^{-2} at 300 nm (dotted line) and 700 nm (full line). Note that the fluence rate at the surface is equal to the irradiance for 300 nm but that the fluence rate is approximately 5.5 times higher than the irradiance at 700 nm. (Results courtesy of Kevin Diamond, Hamilton Regional Cancer Centre.)

For a beam of infinite radius, the slope would be equal to the effective attenuation coefficient defined in equation (14.12), but in this case the slope is about twice as large because of out-scattering from the relatively small beam. Near the surface, the contributions of scattered and unscattered light to the fluence rate can be discerned. The unscattered radiation contribution resembles the fluence rate at 300 nm. The portion due to scattered light results in a fluence rate at the centre of the beam that is greater than the irradiance. This component also results in considerable blurring of the beam edges, even at the tissue surface. These examples illustrate the importance of wavelength in determining the fluence rate in tissue and, as discussed later, the potential therapeutic interaction.

14.5 Therapeutic Laser–Tissue Interactions

14.5.1 Introduction

In the previous sections, the behaviour of photons was affected by cells or tissue but the cells or tissues were not affected. Hence, light could serve a diagnostic function. When light affects cells or tissues then a therapeutic function is possible. This section describes the basics of how light can affect cells and tissues. The term 'tissue' will be used for simplicity to refer to all levels of biological structure, whether organelles, cells, fibres, epithelial layers, vessels or ducts or organs, and only when necessary will specific tissue structure be mentioned.

The laser provides a special form of light that is collimated, monochromatic, coherent and capable of being delivered as very short intense pulses of energy. A conventional light source such as an arc lamp, an incandescent lamp or a light emitting diode can sometimes serve instead of the laser. *Collimation* of laser light is useful because it enables light to be focused into a very small optical fibre for delivery of light to a remote site within the body, or focused by a lens system (such as a microscope) to a very small focus of intense power. A conventional light source can be delivered directly to a tissue without coupling to a fibre. *Monochromicity* of laser light is useful when a particular wavelength is desired. A strong conventional light source can be filtered by a bandpass filter to yield a narrow bandwidth of light centred at the desired wavelength. A *pulsed* laser can provide short bursts of high power. Sometimes a conventional light source can be shuttered or chopped to provide pulses with significant power. The *coherence* of a laser is not a factor in most laser effects, although work continues to explore the possibility. Usually, the power of a laser beam or the energy of a laser exposure, its wavelength and its pulse duration determine the laser effect. A conventional light source can sometimes provide an equivalent effect, although laser sources often are more convenient. For simplicity, the term 'laser' will be used in this chapter although the term 'light' may be equally appropriate.

Light interacts with tissues in several possible ways. However, the first step usually involves the absorption of a photon. There is a special case where the deflection of a photon by a cell or small organelle exerts a pressure on the object which allows a focused laser beam to mechanically 'trap' a cell and to move the object around, the so-called 'optical tweezer'. This special case is not discussed here. Otherwise, the first step of a laser effect is the absorption of a photon by a tissue.

Figure 14.17 illustrates the complexity of a laser–tissue interaction as a cube with axes of mechanism, biological level and time course. The mechanism of the laser–tissue interaction can be chemical, thermal or mechanical in nature. The level of biological structure that absorbs the photon can vary, for example, a laser effect can occur on the organelle, cell or tissue/organ level. The biological response can vary with time. There is an immediate physical effect, a short-term biological response such as an inflammatory response or an immune system response, and a long-term response such as wound healing or scarring. Any particular laser exposure may involve several sub-volumes in this cube of interaction.

14.5.2 Photochemical Effects

A photochemical event requires the absorption of a photon to provide the energy for the subsequent photochemical reaction. For the photochemical reaction to be powered by changes in electronic state, the wavelength of the photon must be in the ultraviolet, visible or near-infrared spectral ranges. For reactions to involve vibrations in molecular bonds, the photon can be in the mid-infrared or far-infrared spectral ranges.

Photochemical reactions are very common. In nature, the absorption of sunlight by chlorophyll in plants yields excited molecular states whose energy is used to bind water and carbon dioxide together as sugars. In industry, light-activated adhesives are used routinely. In medicine, light is used to activate photosensitive drugs that catalyze oxidation of tissues so as to kill cancer cells.

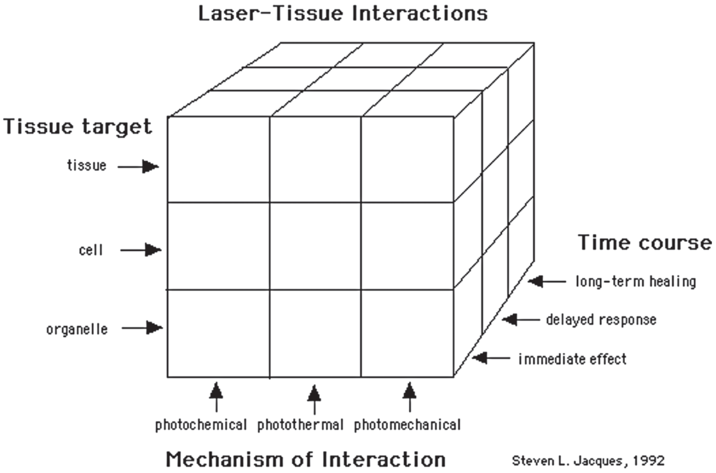


FIGURE 14.17 Cube of laser–tissue interactions. A laser–tissue interaction can involve various mechanisms of interaction, time courses and tissue targets. (Adapted from Jacques 1992 *Surgical Clinics* 72:531–558.)

In photochemical reactions, the photons are a reactant just like the molecules involved. If one delivers an incident irradiance E [W cm^{-2}] to a cuvette filled with non-scattering, negligibly absorbing medium, then the concentration of photons, in units of [mols l^{-1}] or [M], within the cuvette at any moment in time is

$$C_{\mathrm{ph}} = \frac{En}{c}\frac{\lambda}{hc}\frac{1000}{N_{\mathrm{Av}}}. \tag{14.14}$$

The factor En/c is the optical energy density [J cm^{-3}], where n is the refractive index of the medium, c is the *in vacuo* speed of light (3.00×10^{10}cm s^{-1})and c/n is the speed of light in the medium. Multiplication by λ/hc [photons J^{-1}] converts the optical energy density into a photon density [photons cm^{-3}], where h is Planck's constant (6.656×10^{-34}J s). Multiplication by (1000 cm^3 l^{-1})/N_{Av} converts the photon density into a molar concentration of photons [mols l^{-1}], where N_{Av} is Avagadro's number (6.02×10^{23}). For example, an incident irradiance of 1 W cm^{-2} at 488 nm wavelength (an argon ion laser) on a cuvette containing water (n = 1.33) yields a molar photon concentration of 0.18×10^{-12} M.

Assume that these photons are absorbed by a photosensitive molecule with an extinction coefficient s [cm^{-1} M^{-1}] at a concentration C_{ps} [M] in the cuvette. Then the absorption coefficient of the solution containing only photosensitive molecules is $\mu_a = \ln(10)\varepsilon C_{\mathrm{ps}}$, such that transmission T equals $\exp(-\mu_a L)$ where L is the path length of the cuvette. In the literature, ε is reported such that $T = 10^{-\varepsilon CL}$.The rate of absorption of optical energy by the photosensitive molecules is $\mu_a E$ [J cm^{-3}s^{-1}]; therefore, the rate of absorption of photons is $\ln(10)\varepsilon C_{\mathrm{ps}}E(\lambda/hc)\,1000/N_{\mathrm{Av}}$ [M s^{-1}]. But not all absorbed photons may cause a conversion to the photoproduct. There is a quantum efficiency (Φ [dimensionless]) that specifies the fraction of absorbed photons that cause conversion. In summary, the rate of production of a photoproduct, described by its molar concentration C_{pp} [M], is

$$\frac{\partial C_{\mathrm{pp}}}{\partial r} = \ln(10)\varepsilon C_{\mathrm{ps}}E\frac{\lambda}{hc}\frac{1000}{N_{\mathrm{Av}}}\Phi. \tag{14.15}$$

Therefore, the reaction of a concentration of photons (C_{ph}) with a concentration of photoreactive molecules (C_{ps}) to yield a concentration of photoproduct (C_{pp}) is expressed as

$$\text{photon} + \text{photoreactant} \xrightarrow{k} \text{photoproduct}$$

where k [M^{-1}s^{-1}] is the rate constant of the reaction. The rate of production of the photoproduct is

$$\frac{\partial C_{\mathrm{pp}}}{\partial t} = kC_{\mathrm{ph}}C_{\mathrm{ps}}. \tag{14.16}$$

Comparing equations (14.15) and (14.16), the rate of reaction must be

$$k = \frac{\ln(10)\varepsilon c\Phi}{n}. \tag{14.17}$$

The key factor in equation (14.17) is the extinction coefficient ε which describes how efficiently photons are absorbed by the photoreactive molecule.

Photochemical reactions can be treated in the same manner as chemical reactions, with the photons present at a molar concentration proportional to irradiance, E [W cm^{-2}], and reaction rate constants proportional to the extinction coefficient, ε [cm^{-1}M^{-1}], of the photoreactive molecule and to the quantum efficiency of conversion, Φ.

Medical application: Light-activated chemistry is a broad area. Many processes in nature depend on photochemistry, most fundamentally photosynthesis. Light-activated adhesives are well known in industry. Work on developing light-activated adhesives for medicine is in progress. Photochemical treatment of psoriasis of the skin involves light (ultraviolet A) activation of a photoreactant (psoralen) to achieve cross-linking of the DNA in order to control pathological proliferation. PDT is a light-activated chemotherapy. A photosensitive drug is administered to the patient and accumulates in the target tissue such as a cancerous tumour. A treatment light illuminates the tumour to activate the drug. The activated drug passes its activation energy to oxygen to yield oxidizing species that oxidize and kill the tumour. PDT is now being used to treat macular degeneration of the eye (Figure 14.18).

14.5.3 Photothermal Effects

Photothermal reactions are probably the most common form of laser effect used in medicine. The energy of a laser is absorbed by a tissue and converted to heat. Such laser heating can coagulate tissues, similar to the coagulation or 'cooking' of egg white. The coagulation of selected spots on the retina of the eye by an argon ion laser has been a common treatment for a detached retina. The coagulation induces local spots of scarring that holds the retina in place. Laser coagulation of a

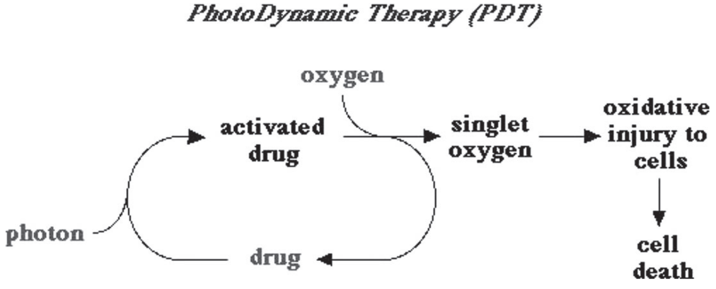


FIGURE 14.18 PDT, a light-activated chemotherapy for treatment of cancer and other disease.

tumour is an effective way to kill a tumour. Vaporization is an effective way to remove a tumour.

14.5.3.1 Laser Heating

The density of laser heating is here referred to as Q [J cm^{-3}] and is described as follows:

$$Q = \mu_a Et \tag{14.18}$$

where μ_a is the absorption coefficient [cm^{-1}], E is the irradiance [W cm^{-2}]and t is the exposure time [s]. The factor Et is often referred to as the radiant exposure, H [J cm^{-2}]. The temperature rise associated with a radiant exposure is

$$\Delta T = \frac{Q}{\rho C_p} \tag{14.19}$$

where ρ is the tissue density [g cm^{-3}] and C_p is the heat capacity [J g^{-1}°C^{-1}]. For water, ρ equals 1 g cm^{-3} and C_p equals 4.18 J g^{-1}°C^{-1}.

Consider a 1 cm^3 cube of water. Deliver a radiant exposure of $H = 1$ J cm^{-2} to one face of the cube (area of face is 1 cm^2). The total energy delivered to the cube is 1 J. If this energy were uniformly absorbed and distributed in the cube, then the temperature rise would be 0.239°C, only a modest temperature rise. Assume that the radiant exposure is at a wavelength where the absorption coefficient for water is μ_a. The energy is deposited as a function of depth z [cm] in the cube as

$$\Delta T(z) = \frac{\mu_a H}{\rho C_p} e^{-\mu_a z}. \tag{14.20}$$

If μ_a is 1 cm^{-1}, then at the surface ΔT equals (1 cm^{-1})(1 J cm^{-2})/(1 g cm^{-3})/(4.18 J g^{-1}°C^{-1}) = 0.239°C. The ΔT drops to 1/e or 37% of this surface value at a depth of $1/\mu_a$ or 1 cm, and this depth is called the *optical penetration depth* δ [cm]. If μ_a is 10 cm^{-1}, then the surface ΔT equals 2.39°C and δ equals 1 mm. If μ_a is 100 cm^{-1}, then the surface ΔT equals 23.9°C and δ equals 100 µm. If μ_a is 1000 cm^{-1}, then the surface ΔT equals 239°C and δ equals 10 µm. So a modest amount of radiant exposure, just 1 J cm^{-2}, that would raise the temperature of a 1 cm^3 cube of water just a fraction of a degree can achieve explosive temperatures if the energy is deposited in a 10 µm thin layer due to strong absorption. This example illustrates how laser heating depends not just on the delivered energy but also on the optical penetration depth and absorption coefficient.

14.5.3.2 Thermal Diffusion

Another factor involved in heating is thermal diffusion. Heat can dissipate from a heated volume by thermal diffusion. A small zone of heating will cool by diffusion more rapidly than a larger zone of heating. The steady-state rise in temperature depends on the rate of energy deposition relative to the size of the zone of energy deposition. Here is an approximate rule of thumb. If the zone of heating is described by a thickness or diameter d [cm], then the time required for thermal relaxation of the heated zone, i.e. the temperature rise dropping to roughly 37% of its initial value, is

$$t_{\text{relax}} = \frac{d^2}{4\alpha} \tag{14.21}$$

where α is the thermal diffusivity [cm^{2}s^{-1}], about 1.4×10^{-3}cm^{2}s^{-1} for water at room temperature. A 1 mm zone of heating will relax in 7.1 s. A 100 µm zone will relax in 0.071 s. A 10 µm zone will relax in 0.00071 s.

A more careful consideration of thermal diffusion makes use of the expression that describes the time-resolved distribution of heat in response to an impulse of heating at time zero:

$$Q(r,t) = Uh(r,t) \tag{14.22}$$

where

$$h(r,t) = \frac{e^{-r^2/(4\alpha t)}}{(4\pi\alpha t)^{m/2}} \tag{14.23}$$

and U is the energy deposited at time zero. The factor $h(r, t)$ is the impulse response. The factor m equals 1 for a 2D sheet of heating and planar diffusion, 2 for a line source of heating and cylindrical diffusion and 3 for a point source of heating and spherical diffusion. r is the distance between the point of observation and the position of initial energy deposition; t is the time of observation. The initial energy deposited is U in units of [J] if $m = 3$, [J cm^{-1} of line source length] if $m = 2$, and [J cm^{-2} of 2D sheet surface area] if $m = 1$. The units of $h(r, t)$ vary with the choice of m:[cm^{-3}] if $m = 3$, [cm^{-2}] if $m = 2$, and [cm^{-1}] if $m = 1$.

The impulse response $h(r, t)$ can be convolved against any initial spatial distribution of heating, $U(r)$. For example, for a 1D problem in which a planar slab of uniform heating, U [J cm^{-2} planar surface], is centred at $z = z_c$ and has a thickness d, the time-resolved response $Q(z, t)$ is:

$$Q(z,t) = \int_{z_c-d/2}^{z_c+d/2} Uh(z - z',t)\,dz' \tag{14.24}$$

where z is the observation point and z' is a local depth of heating within the heated slab. The value of Q (z, t) for each choice of z and t requires evaluating the integral on the right-hand side of equation (14.24). To plot $Q(z, t)$ over all z at one observation time, one must re-evaluate the integral for each choice of z.

Although equation (14.24) looks simple, for very early times the factor t in the denominator of $h(z-z', t)$ is close to zero and so $h(z-z', t)$ increases toward infinity. In practical problems involving a finite value of energy deposition in a finite-size zone of heating, an alternative expression for $h(z-z', t)$ is more useful for computation by a computer algorithm. The $Q(z, t)$ in response to a finite planar zone of thickness d of initial uniform heating U is calculated by equation (14.24) using the following impulse response:

$$h(z,z_c,d,t) = \frac{\sqrt{4\pi\alpha t}}{2(4\pi\alpha t)^{1/2}}\left(\text{erf}\left(\frac{z - z_c + d/2}{\sqrt{4\alpha t}}\right) - \text{erf}\left(\frac{z - z_c - d/2}{\sqrt{4\alpha t}}\right)\right) \tag{14.25}$$

which describes the response observed at position z due to a planar slab of heating centred at depth position z_c with thickness d. In this expression, t can get arbitrarily close to zero, although t cannot equal zero and the behaviour of $h\ (z-z,\ t)$ will properly handle the finite heating zone.

Figure 14.19 compares the response $Q(z,\ t)$ calculated using the $h(z,\ t)$ in equation (14.25) with equation (14.23). The figure illustrates how equation (14.25) appropriately handles the response to a finite slab of uniform heating.

The big advantage of using equation (14.25) is achieved when calculating the response to a non-uniform distribution of initial heating, $U(z)$. Consider the $Q(z,\ t)$ in response to a radiant exposure $H = 1\,\mathrm{J\ cm^{-2}}$ that irradiates the front air/medium surface of a medium with an absorption coefficient $\mu_a = 10\,\mathrm{cm^{-1}}$.The heating as a function of depth is

$$Q(z,t) = \int_0^\infty \mu_a H e^{-\mu_a z'} (h(z,z',\mathrm{d}z',t) + h(z-z',\mathrm{d}z',t))\,\mathrm{d}z' \tag{14.26}$$

where the problem has been discretized in steps of depth dz. The response observed at position z to the energy deposited at time zero in each incremental slab of thickness dz at depth z' is specified, making use of the impulse response ($h(z,\ z',\ \mathrm{d}z',\ t)+h(z,-z',\ \mathrm{d}z',\ t)$) which has two factors. The second factor, $h(z,-z',\ \mathrm{d}z',\ t)$, accounts for an image source that mimics the boundary of thermal insulation at the air/tissue surface. Figure 14.20 illustrates the use of equation (14.26) to calculate the response to the example problem.

14.5.3.3 Thermal Damage

The previous sections have discussed how to calculate the temperature history at some observation point, $T(t) = Q(r,\ t)/(\rho C)$. This section discusses how such a temperature history elicits thermal damage.

Consider a native form molecule (N) that can be converted to a denatured form (D) (see Figure 14.21). There is an energetic barrier to such a conversion. The Arrhenius rate process description of irreversible thermal damage describes the process in terms of a rate (A [s^{-1}]) at which the native form molecule moves to a transition state N* atop the energy barrier and a probability ($\exp(-E/(RT))$ [dimensionless]) that the molecule will convert from the transition state to the denatured state, N* → D. (Here, the symbol E denotes activation energy [J mol^{-1}].) The factor R is the gas constant (8.314 J mol^{-1} K^{-1}). T is temperature expressed in kelvin ($T = 273.15 + T_{\circ C}$). The rate of conversion, k [s^{-1}], from N to D is given by

$$k = A\mathrm{e}^{-E/(RT)}. \tag{14.27}$$

This formalism can be alternatively expressed in terms of the molar entropy, S [J mol^{-1} K^{-1}], and the molar enthalpy,

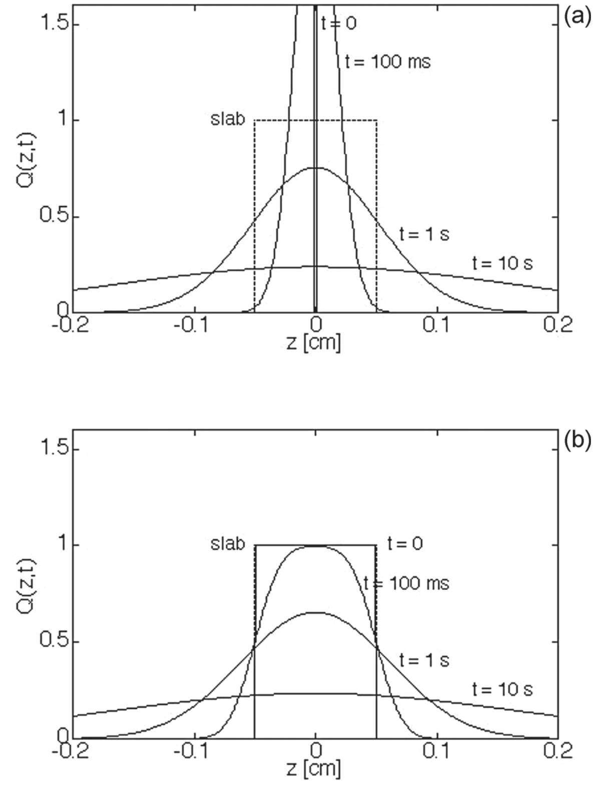


FIGURE 14.19 The spatial distribution, $Q(z,\ t)$ [J cm^{-3}], at four time points in response to a 1 mm thick slab heated by1[J cm^{-2} of planar surface] at time zero. (a) The response using equation (14.23), which does not handle early times. (b) The response using equation (14.25) which properly handles early times. At later times, such as at 10 s, the $Q\ (z,\ t)$ responses calculated by both methods become equal.

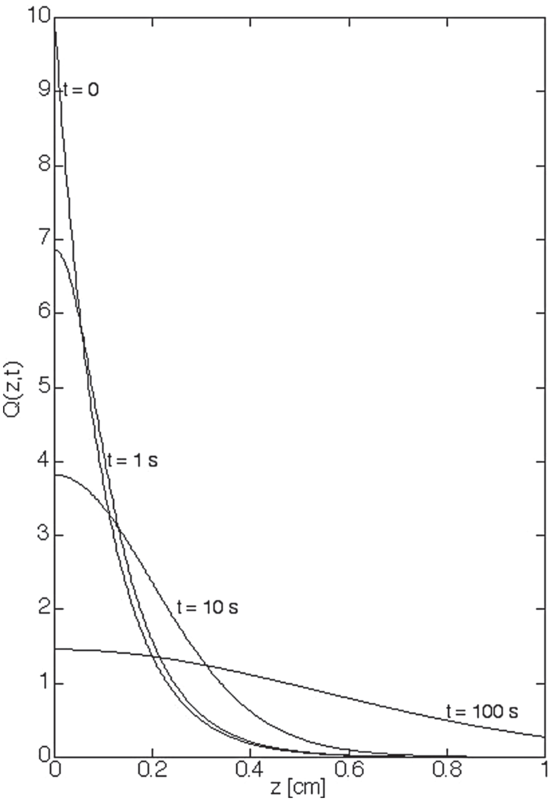


FIGURE 14.20 The spatial distribution, $Q(z,\ t)$ [J cm^{-3}], at four time points in response to irradiating the front air/medium surface of a medium with an absorption coefficient of 1 cm^{-1} by a radiant exposure of 1 [J cm^{-2} of planar surface] at time zero. At time zero, the initial heating is exponentially distributed. At later times, this initial distribution relaxes.

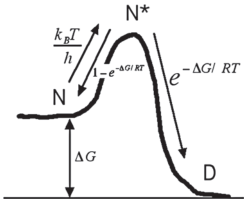


FIGURE 14.21 A native state molecule (N) moves to a transition state (N*) at a rate specified by k_BT/h, then either returns to the native state or converts to an irreversibly denatured form (D).

H [J mol^{-1}]. The free energy, ΔG [J mol^{-1}], associated with the transition from native to denatured form specifies the probability $\exp(^{-}\Delta G/(RT))$. The free energy is related to ΔS and ΔH:

$$\Delta G = \Delta H - T\Delta S \tag{14.28}$$

By convention, the rate at which molecules move from the native state N to the transition state N* is assigned the value k_BT/h, where k_B is the Boltzmann constant (1.381×10^{-23} J K^{-1}) and h is Planck's constant (6.626×10^{-34} J s). This rate is 6.2×10^{12} s^{-1} at room temperature. Therefore, the probability of transition is

$$K = \frac{k_BT}{h}e^{\Delta S/R}e^{-\Delta H/(RT)}. \tag{14.29}$$

The values for A and E that characterize a thermal coagulation process are experimentally specified in experiments in which the rate of coagulation (k) as a function of exposure time and exposure temperature are determined. These values are related to the ΔS and ΔH:

$$A = \frac{k_BT}{h}e^{\Delta S/R} \tag{14.30}$$

$$E = \Delta H.$$

Although the convention that the native molecules sample the transition state N* at a rate k_BT/h is fraught with assumption, the convention suggests that this rate of sampling is proportional to temperature. In contrast, using only A and E to interpret experimental data assumes the sampling is independent of temperature. Note that the value of $\ln(k_BT/h)$ varies over 27.4–29.8 for T in the 40°C–140°C range. This temperature dependence of the sampling frequency is not a critical issue since the values of A are very large and typically experimental data are variable and extend over only a modest temperature range on the Kelvin scale (0°C–100°C equals 273–373 K). So whether one interprets experimental data in terms of A and E or in terms of ΔS and ΔH, one gets about the same answer. More important is the recognition that the temperature dependence of k is dominated by ΔH and that the relatively temperature-independent component of k is affected by ΔS.

If the initial number of native molecules is N_0, then the number of surviving native molecules as a function of time is

$$N(t) = N_0 e^{-\Omega(t)} \tag{14.31}$$

where

$$\Omega(t) = \int_0^t k(t')\,dt'. \tag{14.32}$$

The factor Ω is called the *damage integral* and it extends from zero to infinity, such that $N(t)$ extends from N_0 to zero. At any time, the number of denatured molecules, $D(t)$, equals the difference $N_0 - N(t)$.

A typical experiment might involve wrapping a thin tissue slice in aluminium foil to keep its hydration constant, then immersing the tissue in a hot water bath at a known temperature for a variable amount of time. After the allotted time, the tissue is removed and an experimental measurement is made that is sensitive to the degree of protein aggregation that follows protein denaturation and constitutes 'irreversible coagulation'. There are a variety of transitions that can be monitored by various means to indicate coagulation, for example, loss of collagen birefringence observed through a polarizing microscope, loss of enzyme activity or changes in optical scattering.

Figure 14.22 shows the generic time course for irreversible coagulation for various rate constants k. As the temperature decreases, k decreases and the time required for complete coagulation increases.

The value of k depends on the temperature of exposure and on the values of ΔS and ΔH that characterize some measurable transition in the tissue. There are many such transitions. For example, the irreversible loss of enzymatic activity involves denaturation of labile proteins, the coagulation of egg white and the whitening of soft tissues such as the muscle or liver during cooking involves denaturation and aggregation of proteins into light-scattering aggregates, and the coagulation of collagenous tissue to yield gelatin involves the loss of the collagen fibre bundle structure. The rate constant k at a given temperature can be different for different transitions.

Typical values of ΔS and ΔH for examples of thermal coagulation are listed in Table 14.1 (items 9–12 were calculated from the value of A in table 3.7 in [18], assuming T = 50°C when converting A to ΔS using equation (14.30)). The rate constant k for each example varies as a function of temperature. It is common to consider the exposure time (t_Ω [s]) required to achieve $\Omega = 1$, which corresponds to $t_\Omega = 1/k$, as a function of the temperature. Figure 14.23 plots the exposure time *versus* exposure temperature curves for the several examples of denaturation transitions. The more structured the structure being denatured is, the more vertical the curve becomes as both the change in entropy (ΔS) and the number of bonds to be cooperatively broken (ΔH) are greater. If the molecular structure to be denatured is less structured, the curve is more horizontal as the changes in ΔS and ΔH are less.

The entropy ΔS released by denaturation of a molecular structure involved in an irreversible thermal transition is to a first approximation, linearly related to the enthalpy ΔH of the transition. Figure 14.24a shows this linear relationship. The molar entropy ΔS appears to equal 0.0031 J mol^{-1} K^{-1} per J mol^{-1} of molar enthalpy ΔH. Since hydrogen bonds have a ΔH of about 19 J mol^{-1}, one can estimate how the number of hydrogen bonds cooperatively stabilizing a protein will affect the denaturation kinetics. Figure 14.24b shows the expected

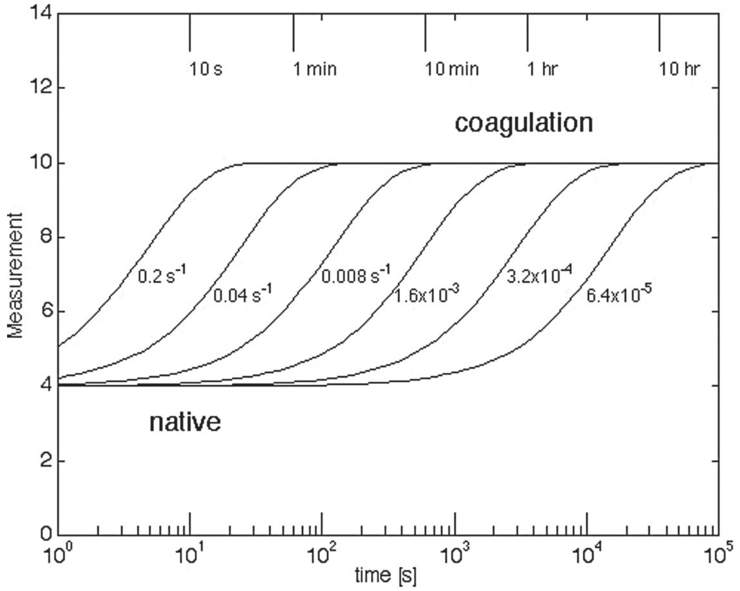


FIGURE 14.22 Time course of irreversible coagulation process for various rate constants, k [s^{-1}]. The x-axis is time of exposure at a constant temperature and, hence, at a constant k. The y-axis is a generic linear measurement of the degree of coagulation (the units are arbitrary, native = 4 and coagulated = 10). As the temperature of the water bath decreases, the value of the rate constant k decreases from 0.2 s^{-1} to 6.4×10^{-5} s^{-1}, and the process of complete coagulation increases from 10 s to 10 h.

TABLE 14.1

Thermal Coagulation Parameters

Label	Type of Necrosis (Reference)	ΔS (J mol^{-1} K^{-1})	ΔH ($\times 10^5$ J mol^{-1})
1	Necrosis of rat liver [19]	392.3	1.866
2	Whitening of dog prostate [20]	277.4	2.596
3	Whitening of dog heart [20]	510.0	2.596
4	Whitening of pig liver [20]	553.5	2.769
5	Contraction of mouse dermis [21]	996.2	4.251
6	Collagen in lens capsule [22]	2381	8.599
7	Loss of rat tail collagen birefringence [23]	831.2	3.676
8	Retinal coagulation [24]	596.7	2.900
9	Retina coagulation [25]	991.7	4.200
10	Retina coagulation [25]	1764.0	6.700
11	Skin coagulation [26]	1640.0	6.300

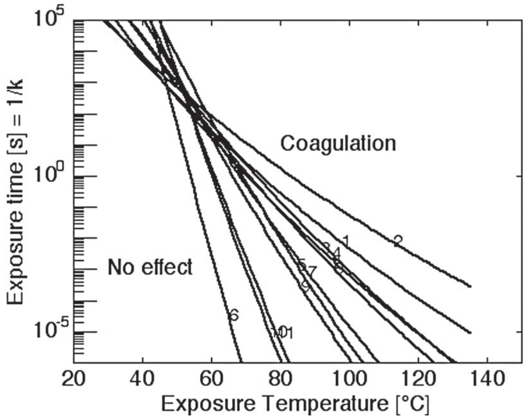
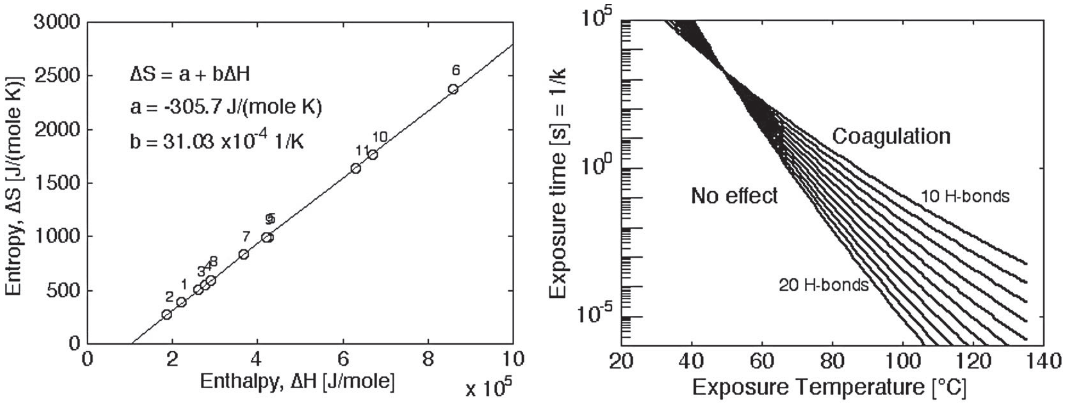


FIGURE 14.23 Exposure time *versus* exposure temperature to achieve thermal damage for various types of transition (e.g. necrosis, whitening, dermal shrinkage, collagen birefringence loss) in various tissues. (Number labels refer to Table 14.1.)

behaviour of exposure time *versus* exposure temperature for a molecular structure held together by increasing numbers of hydrogen bonds, 10–20 bonds. The curves appear to pivot around a point at 49°C for 30 min. Unless all thermal transitions in tissue proceed at the same rate at 49°C, this linear relationship between ΔS and ΔH is an oversimplification and the relationship deserves more investigation. Nevertheless, the relationship suggests that, to a first approximation, only one independent parameter affects coagulation. The choice of ΔH as the independent variable is conceptually connected to the total energy of all bonds that cooperatively hold a molecular structure together. The ΔS is the dependent variable that describes the entropic energy change involved with breaking those bonds.

Medical application: Thermal coagulation of tissue was one of the earliest uses of lasers in medicine. The argon ion laser (488 nm wavelength) is used to coagulate the retina of the eye to cause a local scarring that tacks down a retina that is at risk



(*a*) (*b*)

FIGURE 14.24 Relationship between ΔS and ΔH. (*a*) Plot of molar entropy ΔS *versus* molar enthalpy ΔS for the five examples of Figure 14.23. The ΔS *is* linearly related to ΔH. (Number labels refer to Table 14.1.) (b) Exposure time *versus* exposure temperature based on ΔH, with $\Delta S = a + b\Delta H$, for a molecular structure held together by increasing numbers of hydrogen bonds, 10–20 bonds. Compare with Figure 14.23.

for detachment. Coagulation of small tumours is a common application and can be achieved with various lasers, although a Nd:YAG laser (1064 nm wavelength) will better penetrate tissue and coagulate a larger volume than most other lasers. The body will resorb the coagulated tissue as part of the wound healing process.

A dermatological application of thermal injury is the treatment of vascular lesions such as port-wine stains by a pulsed laser (1–10 ms pulse duration in the 532–585 nm wavelength range, green to orange). The pulsed dye laser deposits energy selectively in the blood vessels due to the strong absorption of light by haemoglobin. The high temperatures transiently achieved in the blood vessels cause thermal damage to the vessel wall and elicit thrombosis (or clotting) of the vessel. The body's wound healing response then removes the clotted vessel, hence removing the port-wine stain lesion.

14.5.4 Photomechanical Effects

When laser energy is deposited in a tissue as thermal energy, there are four types of mechanical effects that may occur:

1. thermoelastic expansion,
2. spallation or cavitation,
3. vaporization and
4. plasma formation.

14.5.4.1 Thermoelastic Expansion

The energy deposition raises the temperature of the tissue by a temperature jump, $\Delta T = \mu_a H/(\rho C_p)$ [°C]. This temperature jump elicits thermoelastic expansion of the tissue expressed as strain (ε) which is the change in length (ΔL) normalized by the initial length (L), $\varepsilon = \Delta L/L$. The strain is related to the temperature jump, $\varepsilon = \beta\Delta T$ [dimensionless], where β is the thermal expansivity [°C^{-1}]. The stress (or pressure) generated by this strain is expressed as $P = M\varepsilon$, where M is the bulk modulus [J m^{-3}] that specifies the pressure per unit strain. In summary, the pressure exerted by thermoelastic expansion is

$$P = M\beta\frac{\mu_a H}{\rho C_p} = \Gamma\mu_a H \tag{14.33}$$

where Γ is a dimensionless parameter called the Grüneisen coefficient that specifies the fraction of thermal energy deposition that couples into mechanical energy. The value of Γ varies with the type of material, the temperature and the rate of heating. The Γ of water at room temperature is about 0.12 for water. About 12% of thermal energy deposited by a 10 ns-pulse laser in biological tissues couples into mechanical energy.

For reference, 1 J m^{-3} of pressure is called 1 Pascal [Pa] and 10^5 Pa equals about 1 bar [bar] or 1 atmosphere. An energy deposition of 1 J cm^{-3} (or 10^6 J m^{-3}) in water at room temperature by a 10 ns-pulse laser would yield 0.12 J cm^{-3} or 1.2×10^5 Pa or 1.2 bar of pressure. This energy deposition corresponds to a temperature jump of only 0.24°C. So modest temperature jumps elicited by pulsed lasers can create significant pressures via thermoelastic expansion.

To achieve this maximum amount of pressure, the laser pulse must deposit its energy in the target tissue volume faster than the speed of sound can dissipate mechanical energy from the volume as pressure waves propagate away from the volume. This concept is referred to as *stress confinement.* The laser pulse duration, t_p [s], required to achieve stress confinement in a tissue volume whose size is characterized by a length d [m] is

$$t_p < \frac{d}{c_s} \tag{14.34}$$

where c_s [m s^{-1}] is the velocity of sound in the tissue. The value of c_s for water is 1480 m s^{-1}, and values for soft tissues are similar. If the laser pulse is significantly longer than d/c_s, the pressure dissipates faster than it can accumulate and the maximum pressure is not achieved.

Once a laser pulse has deposited pressure under stress-confined conditions, the pressure will propagate away from the heated volume at the speed of sound as a pressure wave. The pressure observed at some point is calculated in terms of a related quantity called the *velocity potential,* ϕ [m^{2}s^{-1}], that is proportional to the energy deposition, here called $W = \mu_a H$ [J m^{-3}]. The definition of ϕ is

$$\phi(r,t) = -\frac{\beta}{4\pi\rho C_p} \oiiint_V \frac{W(r')}{|r - r'|} \delta\left(t - \frac{|r - r'|}{c_s}\right) d^3r' \quad (14.35)$$

where $\delta()$ is a Dirac delta function which equals unity only when its argument equals zero, i.e. when $t = |r - r'|/c_s$. This equation states that energy deposited (W) in a small tissue volume at r' which is a distance $c_s t$ from the observation point at r will linearly contribute to the velocity potential observed at location r and time t. However, the contribution falls as the inverse of the distance between source and observation point, i.e. as $1/|r - r'|$ which equals $1/(c_s t)$.

The pressure associated with the velocity potential is expressed as

$$P = -\rho \frac{\partial \phi}{\partial t}. \quad (14.36)$$

Equation (14.22) can be re-stated in a form that allows numerical evaluation of $\phi(r, t)$. If the numerical algorithm uses time steps Δt [s] and the volume of interest where energy has been deposited is divided into sub-volumes, $V[j]$, and the local energy deposition in the jth sub-volume is $W[j]$, and each sub-volume is located a distance $r[j]$ from the observation point, then the value of ϕ observed at time $t = r[j]/c_s$ is

$$\phi[k] = -\frac{\beta}{4\pi\rho C_p} \frac{1}{\Delta t} \sum_j \frac{W[j]}{r[j]} V[j] \quad (14.37)$$

where k is a time index defined as

$$k = \text{round}(r[j]/c_s/\Delta t) \quad (14.38)$$

and time t is

$$t[k] = (k - 0.5)\Delta t. \quad (14.39)$$

Note that by convention, a distribution of energy deposition W generates a time course of negative values of velocity potential ϕ. Think of a ball bouncing off a wall. The force on the wall is positive but the change in velocity of the ball is negative. Therefore, $-\phi$ is a positive factor and is here called the 'negative velocity potential'.

Equations (14.37)–(14.39) will fill arrays $\phi[k]$ and $t[k]$ that specify the time-resolved function $\phi(t)$ at the chosen observation point. Each jth voxel of energy deposition will contribute to a kth time bin of response. The time step Δt and the bin size Δr of the voxels are related as $\Delta t = \Delta r/c$. The pressure is calculated from

$$P[k] = -\rho \frac{\phi[k] - \phi[k-1]}{\Delta t}. \quad (14.40)$$

Figure 14.25 illustrates these calculations for the case of a 4 mm diameter aqueous sphere (density ρ = 1000 kg m^{-3}) at room temperature (Grüneisen coefficient Γ = 0.12) that receives a uniform distribution of energy deposition ($W = 10^6$ J m^{-3}) at time zero. The speed of sound in water, c_s, is 1480 m s^{-1}. The value of β equals $\Gamma C_p/c_s^2$, which equals 2.29×10^{-4}°C^{-1}.

In the first case, the point of observation is at the centre of the sphere. The shell of integration grows as $4\pi r^2$ but the $1/r$ factor in calculating ϕ yields a value for $-\phi$ that grows proportional to $r = c_s t$ and, hence, $-\phi$ is proportional to time. The pressure observed is the time derivative of $-\phi$ and is constant at $P = \Gamma W$ or 1.2×10^5 Pa until the effect from the surface boundary of the sphere arrives at the observation point at a time t = radius/c_s = 0.135×10^{-6} s. At this time, the $-\phi$ drops to zero because there is no energy deposition outside the sphere. The sudden drop to zero causes a negative impulse of tensile pressure. Thereafter, both $-\phi$ and P are zero.

In the second case, the same sphere is observed from a distance of 10 mm. There is a time delay of 5.4×10^{-6} s before the first effect of W in the sphere arrives at the observation point and $-\phi$ becomes non-zero. The $-\phi$ increases with time as the incremental hemispherical shell of integration of W advances to the centre of the sphere and, thereafter, $-\phi$ decreases with time as the shell advances to the rear of the sphere. Consequently, P is initially positive, passes through zero, then becomes negative. Both $-\phi$ and P are zero at times greater than $t = (12 \times 10^{-3}\,\text{m})/c_s = 8.1 \times 10^{-6}$ s.

The pressure P observed at some position will always become first positive and then negative at some point in time. The value of $-\phi$ will always initially increase due to some local W distribution, then decrease again to zero. Hence, the pressure which is proportional to the time derivative of $-\phi$ will always be initially positive, then necessarily negative. The positive pressure is called *compressive* pressure and the negative pressure is called *tensile* pressure.

If there is a boundary of mechanical mismatch, such as where the external medium meets the tissue at the surface boundary, then some or all of the effect of W arriving at the boundary will be reflected according to the reflection coefficient, R_{acoustic} [dimensionless]:

$$R_{\text{acoustic}} = \frac{Z_{\text{external}} - Z_{\text{tissue}}}{Z_{\text{external}} + Z_{\text{tissue}}} \quad (14.41)$$

where Z denotes the mechanical impedance of the external medium or the tissue. For the case of air as the external medium, the value of Z_{external} is negligible relative to the value of Z_{tissue}, so $R_{\text{acoustic}} \approx -1$. In other words, if a positive compressive or 'pushing' wave of pressure arrives at the air/tissue surface, the tissue attempts to expand out into the air but snaps back like a rubber band. This snap back is a *tensile* or 'pulling' pressure and is negative. Such an air/tissue boundary can be treated by the method of images, in which a mirror image of the positive W distribution in the tissue is mapped as a negative W in the region outside the tissue. The $-\phi$ that arrives at a point of observation is based on positive W from within the tissue and negative W from the image outside the tissue.

Figure 14.26 illustrates the propagation of a pressure wave generated in an absorbing medium irradiated with a cylindrical 4 mm diameter flat-field pulsed laser beam. The tissue has an air/tissue surface boundary. The beam delivers a radiant exposure of 0.020 W cm^{-2} (200 W m^{-2}) as an impulse at time zero. The medium has an absorption coefficient of 50 cm^{-1}

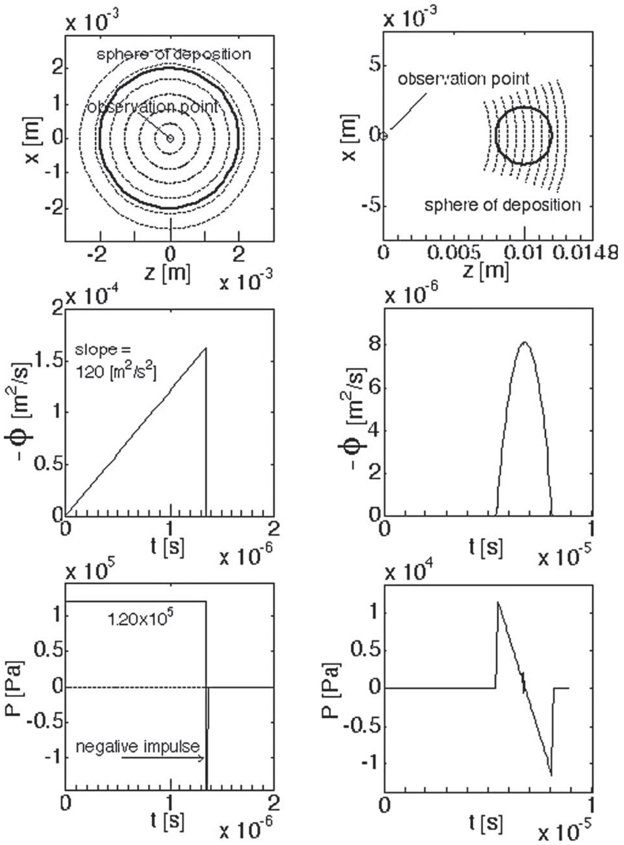


FIGURE 14.25 Time-resolved negative velocity potential ($-\phi$) and pressure (P) generated by an aqueous sphere ($\rho = 1000$ Kg m^{-3}) of uniform energy deposition ($W = 10^6$J m^{-3}) at time zero, at room temperature ($\Gamma = 0.12$): left column, point of observation at centre of sphere; right column, point of observation is a 10 mm distance from sphere; top row, sphere of deposition and point of observation (broken lines indicate shells of integration for calculating $-\phi$ that arrives at observation point at various times); middle row, negative velocity potential, $-\phi$ [m^{2}s^{-1}]; bottom row, pressure, P [Pa].

(5000 m^{-1}). Therefore, the initial W at the surface is 1 J cm^{-3} (10^{6}J m^{-3}), which yields an initial pressure of 1.2×10^5J m^{-3} (assumes $\Gamma = 0.12$ and $\rho = 1000$ kg m^{-3}). The initial W falls off exponentially with depth in the tissue. A negative image of this W distribution is placed above the surface outside the tissue for the calculation. In the central region of the beam, the laser exposure yields a leading positive compressive pressure wave followed by a negative tensile pressure wave reflected from the air/tissue boundary. The discontinuity in W at the edge of the beam also generates a cylindrical negative tensile wave that propagates in toward the central axis as the pressure wave propagates into the tissue. When these inward-travelling tensile waves meet at time $t = \text{radius}/c_s$ a strong negative pressure is created. As the pressure wave propagates into the tissue, these tensile pressures are concentrated into a pulse of negative pressure, sandwiched between two positive pressure pulses. As the wave propagates and the distance from the source grows larger, the front and rear positions of this negative pulse come closer together. The tensile pressure is generated by both the mismatched air/tissue boundary and by the edges of the beam.

Medical application: Thermoelastic expansion itself is not used for therapeutic medical applications but it is the basis for spallation and cavitation discussed in the next section. Such laser-induced pressure waves can be used for diagnostic imaging purposes.

14.5.4.2 Spallation and Cavitation

As tensile pressures develop in a tissue after a laser exposure, there is a limit to the tensile pressure that a tissue can sustain without rupture. When this limit is exceeded, the tissue will 'break' just like a rubber band that has been stretched too far. In solids, such material breakdown involves a process of initial void formation and subsequent void growth usually along a plane of fracture and the process is called *spallation*. In liquids, such a breakage involves the initiation and growth of a bubble and the process is called *cavitation*.

The aqueous nature of most soft tissues makes it likely that cavitation bubbles occur rather than spallation planes. The threshold for cavitation bubbles in water is affected by the purity of the water. For not rigorously pure water, the threshold is about −8 bar or -8×10^5 Pa, which requires an initial temperature jump of 3.2°C($T = -2P/(\Gamma\rho C_p)$ where $\Gamma = 0.12$, $\rho = 1000$ Kg m^{-1}, $C_p = 4184$ J m^3/°C^{-1}, $P = -8\times10^5$ Pa, and the tensile pressure achieves −1/2 the initial maximum compressive pressure at time zero). A very small temperature jump is required to cavitate water. The threshold tensile pressure for cavitation in a 10% collagen gel is about -28×10^{-5} Pa (a 12°C temperature jump) [27]. With increasing mechanical strength, the threshold for cavitation in a medium or tissue increases. Cavitation has been observed photographically in the cornea of the eye when a temperature jump of 40°C was induced by a

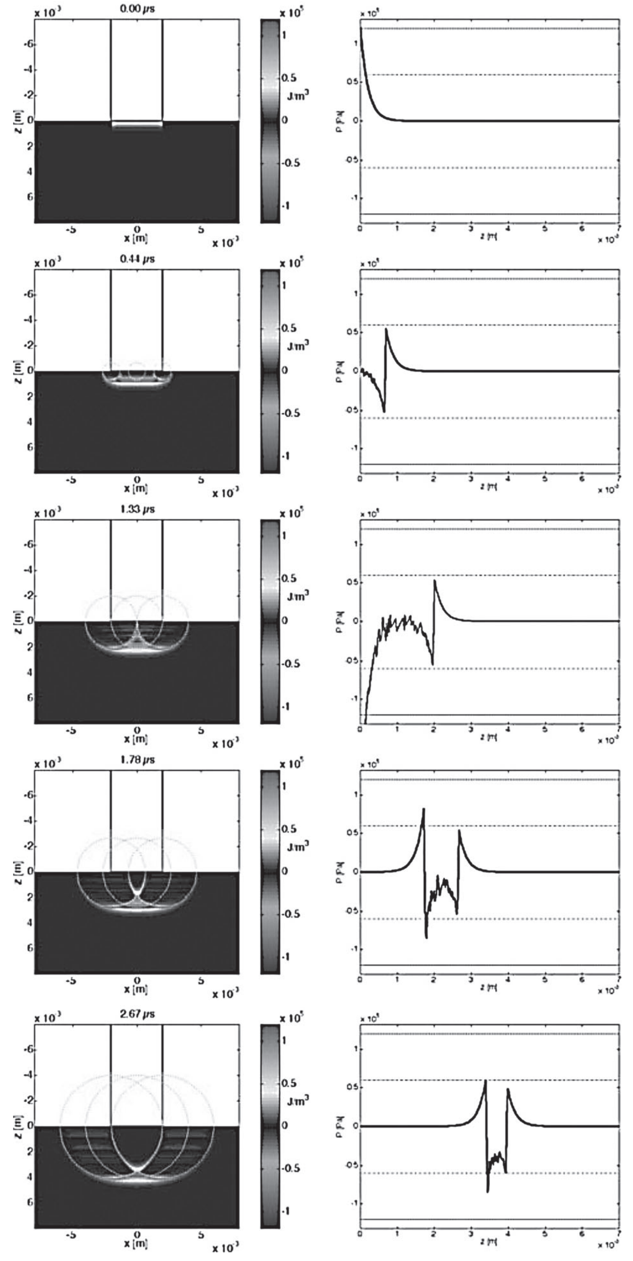


FIGURE 14.26 Time-course of compressive and tensile pressure waves generated by pulsed laser beam penetrating into an absorbing medium with an air/medium surface boundary. Five time points are shown: 0, 0.44, 1.33, 1.78 and 2.67 μs. Left, $P(z, x)$ at each time point. The broken lines indicate spherical shells propagating from the left and right edges and from the centre of beam, to indicate speed of sound. Right, axial $P(z)$ at $x = 0$ at each time. The noise in the traces is an artefact of the numerical simulation.

10 ns pulsed laser [28]. Therefore, it is possible to have cavitation using pulsed-laser-induced temperature jumps that do not cause the tissue temperature to reach 100°C.

Medical application: Spallation and cavitation have not yet been developed into common medical applications. However, such phenomena may enable removal of very thin layers of tissue without significant thermal damage or surface modification of a biomaterial. Investigation into the uses of spallation and cavitation in biomaterials continues.

14.5.4.3 Vaporization

As the temperature of a tissue is raised by a pulsed laser, the water in the tissue can vaporize. The vaporization process can occur in three different ways: surface evaporation, sub-surface explosive vaporization and explosive expansion of superheated fluid. The heat of vaporization of water (Q_{vap}) at 37°C and 100°C is 2.427×10^6 and 2.259×10^6J kg^{-1}, respectively. The apparent molar free energy involved in evaporation at 100°C is $(2.259 \times 10^6 \text{J kg}^{-1})(1 \text{kg l}^{-1})/(55 \text{mol l}^{-1}) = 41.1 \times 10^3 \text{J mol}^{-1}$. This value is comparable to the energy of two hydrogen bonds.

Surface evaporation: If the vapour is formed at the surface, then the vapour evaporates from the surface. Such surface evaporation can desiccate the superficial layers of a tissue. The rate of vaporization is, to a first approximation, equal to the rate of energy deposition in the superficial layers of the tissue. If the laser wavelength is strongly absorbed by tissue such that power is absorbed superficially, then most of that deposited energy is available to drive surface evaporation. For an incident pulsed laser radiant exposure H [J m^{-2}], the evaporation, expressed as the escaping mass of water per unit area of surface M_{evap} [kg m^{-2}], is about

$$M_{evap} = \frac{H}{Q_{vap}}. \tag{14.42}$$

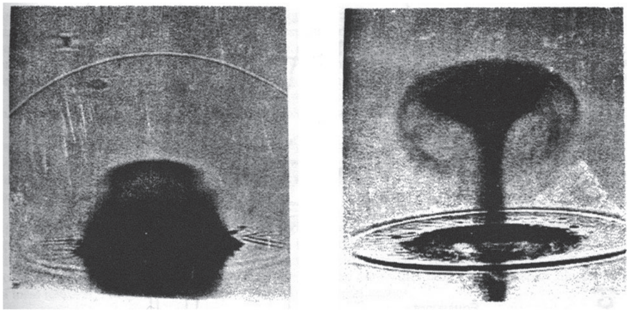
To illustrate desiccation, consider a laser pulse that heats the surface of a tissue from 37°C to 100°C, $\Delta T = (100 - 37) = 63$°C, so there is no explosive vaporization but there is sufficient heat to drive surface evaporation. Let the optical penetration depth of light be 20 μm ($\mu_a = 500 \text{cm}^{-1}$) which is approximately the value of tissue for aCO$_2$ laser (10 600 nm wavelength). The radiant exposure (H) to achieve this heating would be $\Delta T \rho C_p / \mu_a = 0.527 \text{J cm}^{-2}$ or $5.27 \times 10^3 \text{J m}^{-2}$. Then equation (14.29) yields the mass of evaporated water if all the laser energy was used to evaporate water or $(5.27 \times 10^3 \text{Jm}^{-2})/(2.26 \times 10^6 \text{J kg}^{-1}) = 2.33 \times 10^{-3} \text{kg m}^{-2}$ or $2.33 \times 10^{-4} \text{g cm}^{-2}$. Since tissue is basically water, this mass of evaporation would cause the tissue to shrink by a thickness of 2.3 μm.

When a continuous laser irradiates tissue, equation (14.29) can still be used to roughly estimate the rate of vaporization. The laser vaporizes the water and the dry material component of the tissue is carried along by the escaping water. This estimate, however, will often overestimate the amount of tissue removal. There are losses to thermal diffusion which diminish the efficiency. Sometimes carbonization is a key factor in vaporization. For diode lasers in the far red and the Nd:YAG laser in the near infrared, there is only low absorption of laser by the tissue. However, at some point the tissue surface has desiccated and starts to heat up and a carbonized layer forms. At this point, the carbon layer strongly absorbs the laser and the surface becomes quite hot. Surface evaporation proceeds efficiently. The formation of the carbon layer is a cyclic process in which desiccation and heating elicit carbonization, strong absorption by the carbon elicits explosive vaporization, the carbon layer is removed by explosive vaporization, then the cycle repeats. Equation (14.29) can estimate this process by substituting for H by the factor $H\,(1 - T_{carbon})$ where T_{carbon} is the transmission through the average thickness of the carbon layer.

Subsurface explosive vaporization: If the vapour is formed below the surface and cannot evaporate, then the vapour is confined and the pressure builds until it overcomes the strength of the tissue and rapid vapour expansion disrupts the tissue. This has been called the *popcorn effect.* Only a small amount of water needs to vaporize explosively to cause tissue disruption. When the temperature exceeds 100°C by a margin of 10°C–30°C, then there is sufficient energy to allow a small fraction of the tissue water to vaporize. A 10°C–30°C excess temperature in pure water can evaporate 2%–6% of the water. Just a small fraction of the water is converted to vapour and this small amount is sufficient to cause disruptive vaporization in the tissue. For example, the threshold temperature for explosive disruption of red blood cells by a pulsed dye laser during treatment of port-wine stains is about 125°C. The threshold temperature for explosive disruption of melanosomes by a pulsed laser is about 112°C [29].

Superheated fluid expansion: If the temperature rise is sufficient to supply the entire heat of vaporization (Q_{vap}), then the water in the tissue explosively ejects from the tissue. Figure 14.27 shows flash-photography images of the explosive vaporization of water by an Er:YAG laser (2940 nm wavelength). Figure 14.28 shows the distribution of energy deposition that occurred in Figure 14.27. A central disc of water has been supplied with the entire heat of vaporization (here listed at 2654 J cm^{-3}, which would be equivalent to 635°C if temperature did not clamp at 100°C) and it explodes. The region heated to 418–2654 J cm^{-3} (equivalent to 100°C–635°C) does not explode but rather induces rapid surface evaporation that yields a visible annulus of disturbed water.

Medical application: Vaporization is used in medicine to remove bulky tumours, such as obstructing lung tumours or brain tumours. The vaporization process converts the tumour into gas-borne particles that are removed by a vacuum pump.



(*a*) (*b*)

FIGURE 14.27 Explosive vaporization of water by pulsed Er:YAG laser (2940 nm wavelength, 25 mJ, Gaussian beam, 1/e^2 diameter is 3.1 mm). (a) Image at 1.6 μs after pulse. Note shock wave in air. (b) Image at 20 μs. A central 2.4 mm diameter disc of water has been ejected. A larger 3.6 mm diameter annulus undergoes rapid surface evaporation. (Reprinted with permission from SPIE. Jacques and Gofstein 1991 *Proc. SPIE* **1427** 63–7.)

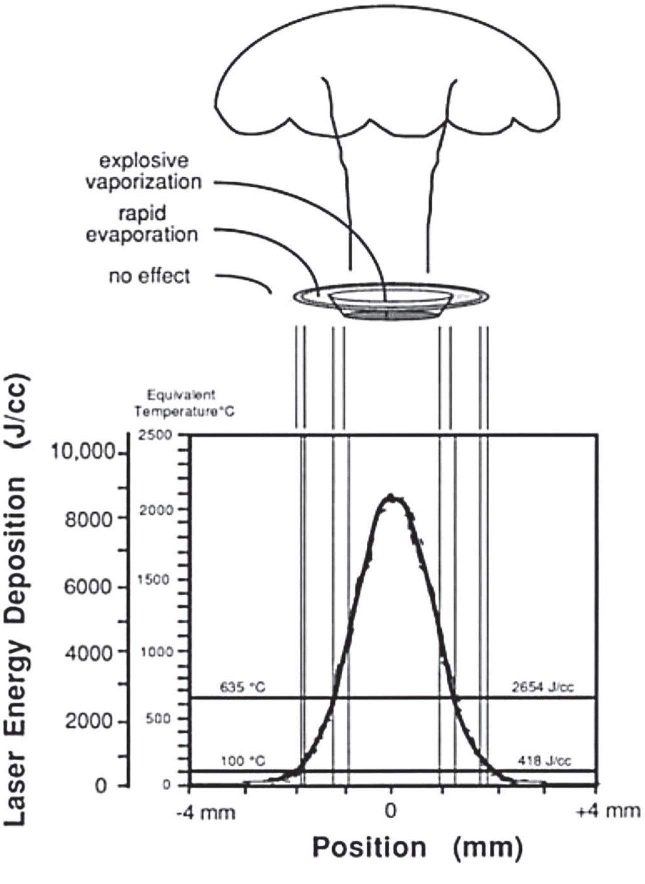


FIGURE 14.28 Spatial distribution of energy deposition for Figure 14.27. The central disc of ejection received 2654 J cm^{-3} of energy deposition that supplied the entire energy of vaporization. The region that received 418–2654 J cm^{-3} did not explode but experienced rapid surface evaporation that appears as a visible annulus. (Reprinted with permission from SPIE. Jacques and Gofstein 1991 *Proc. SPIE* **1427** 63–7.)

Hence, tumours located where access is difficult can be removed by vaporization. The CO_2 laser (10 600 nm wavelength) is strongly absorbed by water and is commonly used for such vaporization. The cw diode laser (e.g. 800 nm wavelength) and Nd:YAG laser (1064 nm wavelength) are often used but rely on desiccation and carbonization of the tumour to provide a black layer to efficiently absorb the laser light.

Another application is the reshaping of the cornea by precise removal of a known thickness of corneal collagen. This process uses the ArF excimer laser (193 nm wavelength) which is very strongly absorbed by protein. The laser removes about μm per laser pulse, and precise control is obtained by controlling the number of pulses delivered. In the common LASIK procedure, a flap of cornea is cut and pulled aside then the laser reshapes the interior surface of the cornea. After completion, the flap is returned to its original position. Hence, the interior rather than the surface of the cornea is removed by the laser to achieve a new corneal shape.

14.5.4.4 Plasma Formation

When the irradiance of a pulsed laser on a tissue exceeds about 1000 W m^{-2} (or 10^7 W cm^{-2}), then optical breakdown can occur and a plasma is formed. This process is complex but essentially involves the ionization of molecules as electrons are stripped by the extremely high electrical fields of the laser light. Focusing a pulsed laser will achieve such high irradiances in the focus. The rapid expansion of the plasma elicits a rapidly expanding bubble that then collapses to produce extreme pressures.

Medical application: Plasma formation has found application in ophthalmology, where a Q-switched Nd:YAG laser (1064 nm wavelength) is focused into the eye near an implanted artificial lens to disrupt a proteinaceous film that interferes with vision. Plasma formation has been used in urology to disrupt kidney stones that are too large to pass normally through the ureter.

For all three cases of cavitation, vaporization and plasma formation, a bubble is often formed that expands to its maximum diameter then collapses to a point and generates very high pressures. This process is very disruptive. The process of bubble collapse is an important mechanism for tissue disruption.

14.6 Summary

Laser–tissue interactions can involve photochemical, photothermal and photomechanical mechanisms of action. The interactions are usually damaging, but controlled local damage can be a therapeutic or surgical tool. Understanding the variety of laser–tissue interactions offers a toolkit of precise local action for medical applications.

REFERENCES

1. Wagnières G. A., Starr W. M. and Wilson B. C. 1998 *In vivo* fluorescence spectroscopy and imaging for oncological applications *Photochem. Photobiol.* **68** 603–32.
2. Springett R., Sakata Y. and Delpy D. T. 2001 Precise measurement of cerebral blood flow in newborn piglets from the bolus passage of indocyanine green *Phys. Med. Biol.* **46** 2209–25.
3. Lakowicz J. R. 1999 *Principles of Fluorescence Spectroscopy* 2nd edn (New York: Kluwer).
4. Turro N J 1978 *Modern Molecular Photochemistry* (Menlo Park, CA: Benjamin/Cummings).
5. Vanderkooi J. M., Maniara G., Green T. J. and Wilson D. F. 1987 An optical method for measurement of dioxygen concentration based upon quenching of phosphorescence *J. Biol. Chem.* **262** 5476–82.
6. Bohren C. F. and Huffman D. R. 1983 *Absorption and Scattering of Light by Small Particles* (New York: Wiley).
7. Stern 1975 In vivo evaluation of microcirculation by coherent light scattering *Nature* **254** 56–8.
8. Hanlon E. B., Manoharan R., Koo T. W., Shafer K. E., Motz J. T., Fitzmaurice M., Kramer J. R., Itzkan I., Dasari R. R. and Feld M. S. 2000 Prospects for *in vivo* Raman spectroscopy *Phys. Med. Biol.* **45** R1–59
9. Sankaran V., Walsh J. T. and Maitland D. J. 2002 Comparative study of polarized light propagation in biologic tissues *J. Biomed. Opt.* **7** 300–6.
10. Baba J. S., Cameron B. D., Theru S. and Cote G. L. 2002 Effect of temperature, pH, and corneal birefringence on polarimetric glucose monitoring in the eye *J. Biomed. Opt.* **7** 321–8.
11. Cheong W-F, Prahl S A and Welch A J 1990 A review of the optical properties of biological tissues *IEEE J. Quantum Electron.* **26** 2166–84
12. Schmitt J. M. and Kumar G. 1998 Optical scattering properties of soft tissue: a discrete particle model *Appl. Opt.* **37** 2788–97.
13. Flock S. T., Wilson B. C. and Patterson M. S. 1987 Total attenuation coefficients and scattering phase functions of tissues and phantom materials at 633 nanometres *Med. Phys.* **14** 835–41.
14. Schmitt J. M. 1999 Optical coherence tomography (OCT): a review *IEEE J. Sel. Top. Quantum Electron.* **5** 1205–15.
15. Beek J. F., Blokland P., Posthumus P., Aalders M., Pickering J. W., Sterenborg H. J. C. M. and van Gemert M. J. C. 1997 In vitro double-integrating-sphere optical properties of tissues between 630 and 1064 nm *Phys. Med. Biol.* **42** 2255–61.
16. Hapke B. 1993 *Theory of Emittance and Reflectance Spectroscopy* (Cambridge: Cambridge University Press).
17. Kalos M. H. and Whitlock P. A. 1986 *Monte Carlo Methods* (New York: Wiley).
18. Niemz M. 2002 *Laser-Tissue Interactions* 2nd edn (Berlin: Springer) p. 79.
19. Matthewson K., Coleridge-Smith P., O'Sullivan J. P., Northfield T. C. and Bown S. G. 1987 Biological effects of intrahepatic neodymium:yttrium-aluminum-garnet laser photocoagulation in rats *Gastroenterology* **93** 550–7.
20. Jacques S. L. 2003 Personal communication of unpublished studies.
21. Jacques S. L. and Prahl S. A. 1987 Modelling optical and thermal distributions in tissue during laser irradiation *Lasers Surg. Med.* **6** 494–503.

22. Miles C. A. 1993 Kinetics of collagen denaturation in mammalian lens capsules studied by differential scanning calorimetry *Int. J. Biol. Macromol.* **15** 265–71.
23. Maitland D. J. and Walsh J. T. Jr 1997 Quantitative measurements of linear birefringence during heating of native collagen *Lasers Surg. Med.* **20** 310–18.
24. Vassiliadis A., Christian H. C. and Dedrick K. G. 1971 Ocular laser threshold investigations *Aerospace Med. Report* F41609-70-0002.
25. Takata A. N., Zaneveld L. and Richter W. 1977 Laser-induced thermal damage in skin *Aerospace Med. Report* SAM-TR-77-38.
26. Henriques F. C. 1947 Studies of thermal injury. V: The predictability and the significance of thermally induced rate processes leading to irreversible epidermal injury *Am. J. Pathol.* **23** 489–502.
27. Esenaliev R. O. and Jacques S. L. 1996 Transient and permanent cavitation in collagen gels by laser-induced thermoelastic pressure waves *Laser–Tissue Interaction (Proc. SPIE 2681)* vol VII (Bellingham, WA: SPIE).
28. Oraevsky A. A., Esenaliev R., Jacques S. L. and Tittel F. K. 1995 Laser flash photography of cold cavitation-driven ablation in tissues *Proc. SPIE* **2391** Laser-Tissue Interaction VI (publ. SPIE)
29. Jacques S. L. and McAuliffe D. J. 1991 The melanosome: threshold temperature for explosive vaporization and internal absorption coefficient during pulsed laser irradiation *Photochem. Photobiol.* **53** 769–75.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

15

Ophthalmic Laser Therapy and Surgery

Daniel Palanker

CONTENTS

15.1 Introduction 227
  15.1.1 Early History 228
  15.1.2 Optical Properties of the Eye 228
15.2 Photothermal Therapy 228
  15.2.1 Photothermal Interactions 228
  15.2.2 Quantification of Thermal Damage 229
  15.2.3 Photocoagulation 229
    15.2.3.1 Retinal Plasticity Following Photocoagulation 230
    15.2.3.2 Optimization of Pulse Duration for Photocoagulation 230
    15.2.3.3 Pattern-Scanning Retinal Photocoagulation 232
    15.2.3.4 Non-damaging Laser Therapy of the Macula 232
    15.2.3.5 Real-Time Monitoring of Tissue Temperature 234
    15.2.3.6 Laser Trabeculoplasty 234
15.3 Tissue-Selective Therapy Using Photochemical Interactions 234
15.4 Photomechanical Interactions 235
  15.4.1 Selective RPE Therapy 235
  15.4.2 Selective Laser Trabeculoplasty 235
  15.4.3 Corneal Ablation for Refractive Surgery 236
15.5 Transparent Tissue Surgery with Ultrashort-Pulse Lasers 237
  15.5.1 Refractive Surgery 237
  15.5.2 Vitreoretinal Surgery 237
  15.5.3 Cataract Surgery 237
15.6 Summary and Future Directions 239
Disclosure Statement 239
References 240

15.1 Introduction

Ophthalmology was the first medical specialty to adopt lasers right after their invention more than 50 years ago, and they gradually revolutionized ocular imaging, diagnostics, therapy, and surgery. Lasers enable exquisitely precise spatial and temporal control of the light delivery to tissue, which can enact a wide variety of interaction mechanisms, including photochemical, photo-thermal, photo-mechanical, and photo-ionization. A broad range of wavelengths and pulse durations provide excellent means for very selective interactions with the target tissue. Seconds-long exposures at low power can activate photosensitizers in photodynamic therapy (PDT), while avoiding tissue heating. Millisecond exposures are used for coagulation of photoreceptors and sealing blood vessels. Below photocoagulation threshold, millisecond exposures can produce sub-lethal hyperthermia, enabling non-damaging therapy of the macula. Microsecond pulses can selectively destroy pigmented cells by exploding melanosomes, without damaging adjacent tissue. Nanosecond-to-femtosecond pulses with high peak irradiance can ionize molecules, enabling dissection of transparent tissues.

Challenging precision, safety, and selectivity requirements for ocular therapeutic and surgical procedures keep advancing the laser technologies, which in turn continue enabling novel applications for the preservation and restoration of sight. Modern lasers can provide single-cell layer selectivity in therapy, sub-micrometre precision in three-dimensional image-guided surgery, and non-damaging retinal therapy under opto-acoustic temperature control. This chapter reviews the evolution of laser technologies; progress in understanding of the laser-tissue interactions; and concepts, misconceptions, and accidental discoveries that led to modern therapeutic and surgical applications of lasers in ophthalmology. It begins with a brief historical overview, followed by a description of the laser–tissue interactions and corresponding ophthalmic applications.

15.1.1 Early History

The concept of ocular phototherapy was first publicized by Dr. Gerhard Meyer-Schwickerath, who took patients to the roof of his laboratory in 1949 and focused sunlight on their retinas to treat melanomas. Because the technique was hardly practical, non-solar sources were sought. Xenon arc photocoagulator, which became available by the mid-1950s, was effective for sealing retinal breaks and treating tumours but was hard to use, and the burns were large and severe.

The development of **laser** in 1960 revolutionized precision and control of light delivery because of several key distinctions from the non-coherent sources: (i) photons are emitted at the same phase (and therefore are coherent); (ii) wavelength range is very narrow (making it monochromatic); and (iii) the beam is very directional (it is well collimated), hence can be tightly focused. Recognition of laser's potential for ocular therapy was almost immediate—the first report on making ocular lesions with a ruby laser was published the next year, in 1961 [1]. The discovery of the argon laser in 1964 provided blue (488-nm) and green (514-nm) wavelengths, which had the advantage of being strongly absorbed by haemoglobin and melanin. Studies on the retinal application of the argon laser were soon underway, with several delivery systems including biomicroscopy [2]. Further progress was achieved when the laser was coupled to a slit lamp, providing an aiming beam and precise control of the spot size, location, power, and exposure duration. A series of clinical papers in 1971 signalled both the dissemination and acceptance of the new technology, which, in various iterations, remained the basis for clinical photocoagulation for the next 50 years.

Continuous progress in understanding of the laser–tissue interactions and advancements in laser technologies, together with discoveries of various tissue response pathways, keep advancing the precision and selectivity of the ocular laser therapy and surgery, and enable development of novel treatment strategies, which are reviewed below.

15.1.2 Optical Properties of the Eye

The relaxed eye has an approximate optical power of 60 Diopters (i.e., its focal length is 16.7 mm in air), with the corneal power being about 40 D, or two-thirds of the total power. Due to orderly arrangement of collagen fibrils in the cornea it is highly transparent, with transmission above 95% in the spectral range of 400–900 nm [3]. The refractive index of the cornea is $n \approx 1.3765 \pm 0.0005$ [3]. The amount of light reaching the retina is regulated by the pupil size, which can vary between 1.5 and 8 mm. The anterior chamber of the eye, which is located between the cornea and lens capsule, is filled with a clear liquid—the aqueous humour, which has a refractive index of $n \approx 1.3335$. The crystalline lens of the eye, located behind the iris, is composed of specialized crystalline proteins with refractive index of $n = 1.40$–1.42. The lens is about 4 mm in thickness and 10 mm in diameter and is enclosed in a tough, thin (5–15 μm), transparent collagenous capsule. In the relaxed eye, the lens has a power of about 20 D, while in the fully accommodated state it can temporarily increase up to 33 D. The vitreous humour, a transparent jelly-like substance filling the large cavity posterior to the lens, and anterior to the retina, has a refractive index of $n \approx 1.335$ [3].

Light entering the eye can be reflected, scattered, transmitted, or absorbed. Reflected or scattered light contains information that can be used for non-invasive diagnostic purposes. The absorption characteristics of ocular tissues are determined by the chromophores resident within the tissue. In the visible part of the spectrum (400–800 nm) these chromophores include: (i) melanin located in the retinal and iris pigmented epithelia, choroid, uvea, and trabecular meshwork (TM); (ii) haemoglobin, located in the red blood cells; (iii) macular xanthophyll, which is located in the plexiform layers of the retina, especially near and in the macula; (iv) rhodopsin and cone photopigments which are located in the photoreceptors; and (v) lipofuscin, located primarily in the RPE layer. These pigments are of major importance in absorption of visible light in the retina from both physiologic and pathologic standpoints. The absorption spectra of these pigments, as well as water and proteins, are illustrated in Figure 15.1a. In the mid-infrared part of the spectrum (3–15 μm) the major absorber is water, while in ultraviolet (below 250 nm), protein absorption is dominant. Absorption of the retinal layers at 532nm wavelength is illustrated in Figure 15.1b.

15.2 Photothermal Therapy

15.2.1 Photothermal Interactions

Absorption of light in tissue leads to heating, and temperature is the governing parameter in all thermal laser–tissue interactions. Tissue effects depend on temperature and exposure duration, and may include sub-lethal response to hyperthermia, as well as necrosis, coagulation, vaporization, and carbonization.

In addition to the laser irradiance, I and absorption coefficient μ_a, which depends on the laser wavelength, heat generation in tissue is determined by the exposure time. If heat diffusion is neglected, then at a constant beam intensity the temperature rise T is linear with time: $T(z,t) = \dfrac{\mu_a \cdot I(z) \cdot t}{\rho \cdot c}$, where z is a depth coordinate, ρ is tissue density, and c is its heat capacity ($c = 4.2$ J g^{-1} K^{-1}, $\rho = 1$ g cm^{-3} for water). Heat diffusion significantly reduces the temperature rise if pulse duration exceeds the characteristic time necessary for heat to spread beyond the zone of light penetration in tissue. For the laser penetration depth L, the heat diffusion time is $\tau = L^2/4\alpha$, where α is thermal diffusivity ($\alpha = 1.4 \times 10^{-3}$ cm^2 s^{-1} for water). For example, for $L = 1$ μm in water, the characteristic heat diffusion time is $t = 1.7$ μs, whereas for $L = 1$ mm, the diffusion time $t = 1.7$ s. If pulse duration is comparable to or exceeds the characteristic diffusion time across the light absorption zone, then proper estimation of the peak temperature in tissue should take heat diffusion into account.

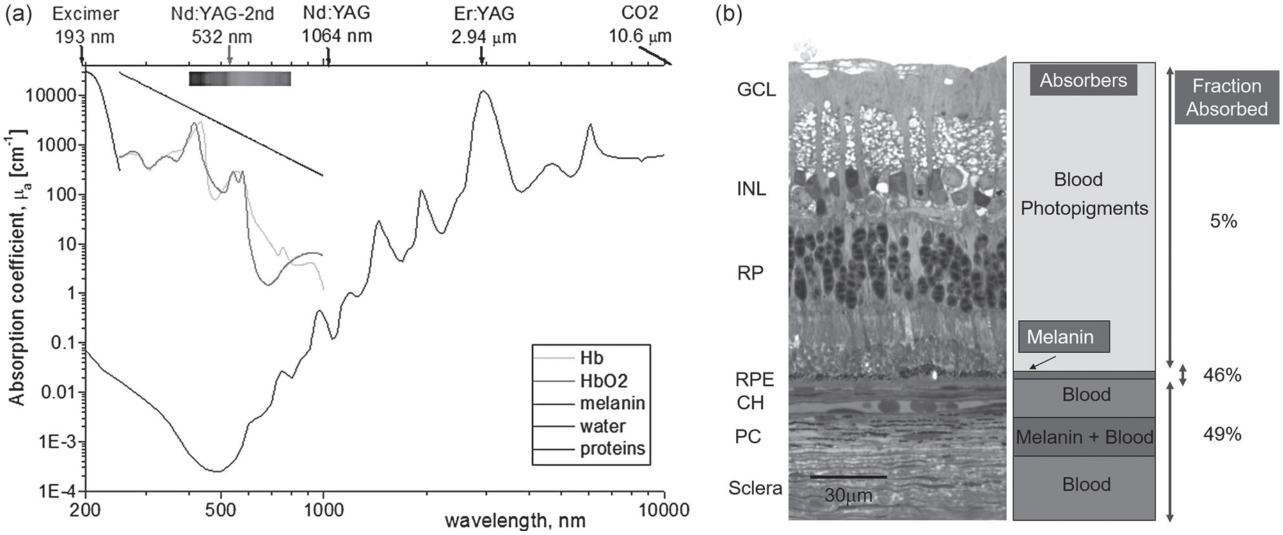


FIGURE 15.1 (a) Absorption coefficients for major chromophores in a spectral range of 0.2–10 μm. Spectral locations of some of the popular laser lines in this range are indicated above the plots. (Adapted from Ref. [4].) (b) Rabbit retina, its major chromophores, and relative fraction of the 532 nm light absorption in its layers. Abbreviations: GCL – ganglion cell layer, INL – inner nuclear layer, PR – photoreceptors, RPE – retinal pigmented epithelium, CH – choroid, PC – pigmented choroid.

15.2.2 Quantification of Thermal Damage

Temperature rise induces abnormal conformational changes in proteins and other biomolecules, which denature at characteristic rates specific to molecular species. Depending on the temperature and duration of the hyperthermia, these thermal processes may eventually lead to cell necrosis. Thermal damage to cells in a millisecond range of pulse durations can be approximated using the Arrhenius model [5] of thermal denaturation. It assumes a rate of decline in the concentration of a critical molecular component $D(t)$ with temperature $T(t)$:

$$\mathrm{d}D(t) = -D(t)\cdot A\cdot \exp\left(-\frac{E^*}{R\cdot T(t)}\right)\mathrm{d}t, \qquad (15.1)$$

where E^* and A are the activation energy and rate constant parameterizing the process, and $R = 8.3\ \mathrm{J}\ (\mathrm{K}\cdot\mathrm{mol})^{-1}$ is the gas constant. Tissue damage, expressed as a decrease in critical molecular component $D(\tau)$, relative to its initial value D_0 over the pulse length τ, is represented by the Arrhenius integral Ω as follows:

$$\Omega(\tau) = -\ln\left(\frac{D(\tau)}{D_0}\right) = A\int_0^{\tau}\exp\left(-\frac{E^*}{R\cdot T(t)}\right)\mathrm{d}t. \qquad (15.2)$$

The model assumes that irreversible tissue damage occurs when concentration of the critical molecular component drops below some threshold value. For example, if this threshold corresponds to a reduction in concentration by a factor of e, then $\Omega = 1$, and the remaining fraction of intact protein is $1/e \approx 37\%$—or in other words, 63% of a certain protein has been damaged.

Measurements of the retinal pigmented epithelium (RPE) damage at various irradiation conditions yield the following average values [5]: $E^* = 340\ \mathrm{kJ\ mol^{-1}}$ and $A = 1.6\times10^{55}\ \mathrm{s^{-1}}$. It is important to keep in mind though that accurate estimation of cell survival under thermal stress is much more complex than just assessment of the denaturation rate of one type of protein or another. For example, (i) multiple types of proteins denature at different rates, (ii) protein significance for cellular survival varies, and (iii) cellular repair mechanisms cannot be ignored at long exposures. Therefore, the single values of the reaction rate A and activation energy E^* most likely represent characteristics of the so-called weak link in cellular metabolism—the most susceptible molecule to thermal damage—and may vary in different cell types.

The Arrhenius model fails to predict correct threshold temperatures at exposures longer than approximately 1 s because cellular repair mechanisms start playing a role at long exposures [5]. At pulse durations shorter than 50 μs and temperatures below the vaporization threshold (~100°C), the coagulation is insufficient for lethal damage, and therefore, the dominant mechanism of damage at shorter pulses becomes mechanical: cell rupture by vapour bubbles forming around overheated melanosomes [6].

15.2.3 Photocoagulation

Retinal photocoagulation typically involves application of laser pulses with durations ranging from 10 to 200 ms and transient temperature rise by tens of degrees. Laser energy is absorbed primarily by melanin in the RPE and choroid, and by haemoglobin in blood. Currently, the most common lasers in photocoagulation are frequency-doubled Nd:YAG (532 nm) and yellow semiconductor laser (577 nm). Yellow wavelength has a slightly lower absorption coefficient in melanin than green has, but lower scattering of longer wavelength makes it less dependent on scattering in edematous retina or on other ocular opacities, which improves consistency and reproducibility of the photocoagulation [7]. At a 532-nm wavelength, only 5% of the incident laser energy is absorbed in the neural retina, approximately 45% in the RPE, and the rest in the choroid (Figure 15.1b) [5]. The generated heat diffuses from the RPE and choroid into the retina and causes coagulation of the photoreceptors and, sometimes, of the inner retina. During 100-ms applications, the heat diffuses by up to 200 μm, thus "smoothing" the edge and extending the coagulated zone beyond the boundaries of the laser spot; this effect is termed as thermal blooming. When one uses shorter pulses and smaller spot sizes, heat diffusion can be limited to the photoreceptor layer (approximately 70 μm in thickness), thereby avoiding the inner retinal damage.

Photocoagulation was proven safe and effective in treatment of proliferative diabetic retinopathy. In this disorder, the retina becomes ischemic and releases a variety of chemical messengers, most importantly vascular endothelial growth factor (VEGF), and these cytokines stimulate the growth of new blood vessels and increase retinal vascular permeability. The abnormal new vessels, associated fibrous tissue, and macular edema are major causes of the sight-threatening complications in the diabetic eye disease. The premise for initial introduction of lasers to treatment of proliferative diabetic retinopathy in 1969 [8] was based on observation that patients with retinal scars (which lack photoreceptors) do not develop neovascularization. Beetham, Aiello, and co-authors (1970) then suggested that creating retinal scars by laser photocoagulation may help prevent the proliferative phase of diabetic retinopathy. This premise now appears erroneous, and just elimination of a significant fraction (approximately 30%) of photoreceptors (the metabolically most active and most numerous cells in the retina) reduces metabolic demand and thereby decreases the production of angiogenic factors. Scarring appears to be a secondary phenomenon associated with the loss of photoreceptors but is not a requirement. In fact, it can be minimized in photocoagulation, as is discussed below.

Coagulation of a large fraction of photoreceptors (approximately 30%) in the peripheral retina (Figure 15.2) has been termed panretinal photocoagulation (PRP), and it significantly reduces the risk of central vision loss due to neovascularization [9,10]. The side effects of PRP—mild nyctalopia (difficulty seeing at low light) and reduction of the peripheral visual field—are outweighed by preservation of the central vision.

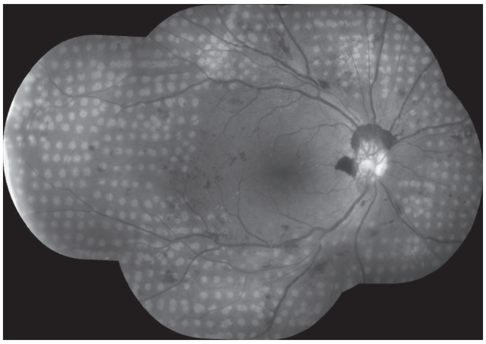


FIGURE 15.2 Fundus of a patient with diabetic retinopathy after PRP. All the peripheral retinas beyond the vascular arcades are treated with lesions (light round spots) coagulating photoreceptors and RPE, but preserving the inner retina. Elimination of up to 30% of the photoreceptors, the metabolically most active and numerous cells in the retina, reduces oxygen consumption and thereby decreases angiogenic signalling, preventing neovascularization and sparing the central vision. (Figure adapted from Ref. [11].)

15.2.3.1 Retinal Plasticity Following Photocoagulation

Initially, relatively long pulses (100–200 ms) have been used to produce rather intense lesions in retinal coagulation [9]. Effects of an intense burn in a retina produced by a 100-ms laser pulse include acute full thickness injury, shown Figure 15.3a, resulting in permanent scarring, shown in Figure 15.3b. Such intense lesions are unacceptable because they destroy not only photoreceptors but also the nerve fibre layer, thereby creating large arc defects in the visual field. More conventional lesions at present are the so-called light burns, illustrated in Figure 15.3c and d. Photoreceptors are damaged, but the inner nuclear layer and ganglion cells are very well preserved, although light burns also result in permanent scarring, as shown in Figure 15.3d. In intense burns, these scars tend to expand over time, thereby enlarging the scotomata [12].

Interestingly, in small and very light lesions (minimally visible or barely visible clinical grades), the glial scar forming in place of the destroyed photoreceptors contracts over time, and photoreceptors shift from adjacent areas into the lesion, as shown in Figure 15.4. In this process, the migrating photoreceptors rewire to the local differentiated bipolar cells, restoring on and off retinal signal pathways in the lesion [13]. This process allows achievement of the goal of PRP—reduction of the retinal metabolic load by destroying a significant fraction of photoreceptors (approximately 30%), metabolically the most active and the most numerous cells in the retina, while avoiding the deleterious side effects of scotomata and scarring, which degrade peripheral vision.

15.2.3.2 Optimization of Pulse Duration for Photocoagulation

Owing to heat diffusion, lesion diameter increases with pulse duration. Therefore, the more confined and better-healing lesions can be achieved using shorter pulses. Another benefit of shorter exposures is the reduced sensation of pain [14] owing to decreased diffusion of heat into the choroid, where pain receptors are located.

Because the Arrhenius integral scales linearly with pulse duration (equation 15.2), higher temperatures are required to achieve the same clinical grade with shorter pulses. In addition, higher power is required to deliver the same energy with shorter pulses. If temperature exceeds the vaporization threshold, transient vapour bubble may result in retinal rupture. As illustrated in Figure 15.5a, for pulse durations of 20, 50, and 100 ms, all the grades (light, moderate, intense, very intense) could be created with appropriate choice of power settings, without retinal rupture. At pulse durations below 10 ms, creating intense lesions without inadvertently rupturing the retina becomes increasingly difficult. At 1 ms, there was practically no difference between the power required to create a mild retinal lesion and that required to produce a rupture [15].

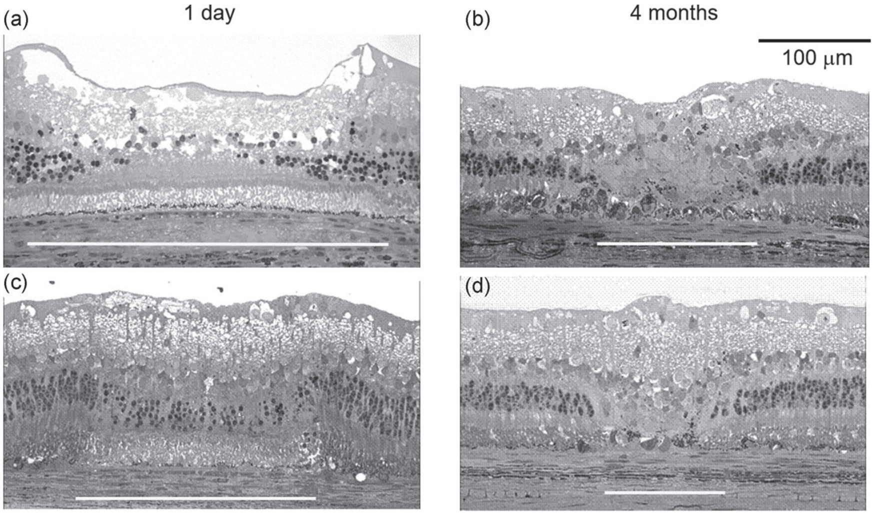


FIGURE 15.3 (a) An intense burn in a rabbit retina produced by 100-ms laser pulse, including full thickness injury and early necrotic features 1 day after treatment, and (b) 4 months after treatment. (c) Light burn exhibits damaged photoreceptors at 1 day, but the inner nuclear layer and ganglion cells are preserved. (d) Residual permanent scarring of the light burn at 4 months. Large scars tend to slowly expand over time (so-called scar creep), further extending the scotoma. (Figure adapted from Ref. [11].)

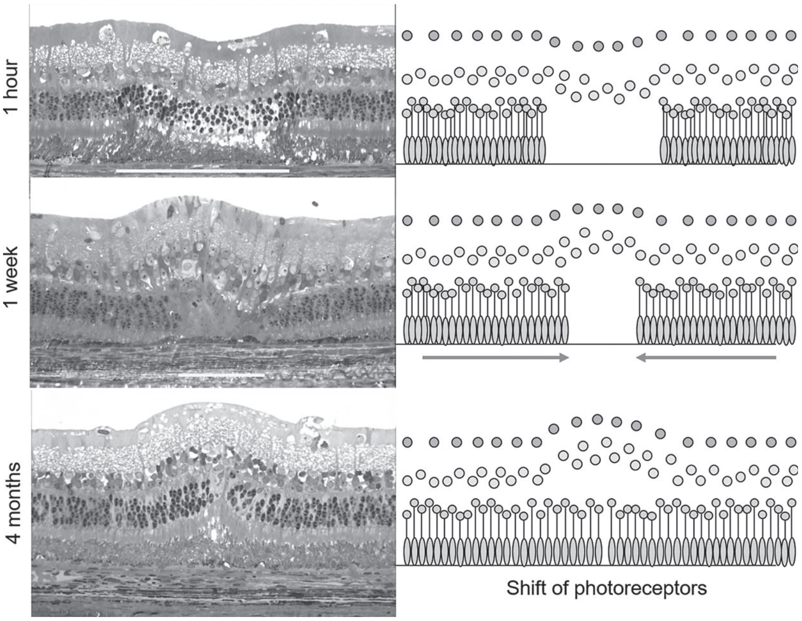


FIGURE 15.4 Migration of the photoreceptors into the very light lesion of 200 μm in width restores continuity of the photoreceptor layer over time and prevents formation of a scotoma and scarring. Shifted rods and cones rewire to the local bipolar cells, restoring retinal neural network in the lesion, including proper function of the on and off pathways. (Figure adapted from Ref. [11].)

The therapeutic window, or a safe dynamic range of retinal photocoagulation, can be quantified as a ratio of the threshold power producing a rupture to that required for producing a light lesion. The larger this ratio, the greater is the margin of safety to create a visible lesion without inadvertently inducing a retinal rupture. Figure 15.5b depicts the width of this therapeutic window as a function of pulse duration for two different laser spot sizes. As pulse duration decreases from 100 to 1 ms, the width of the therapeutic window declines from 3.9 or 5.4 to 1, for 132-μm or 330-μm retinal spot sizes, respectively. The width of the safe therapeutic window should suffice to accommodate for variations in fundus pigmentation, which typically do not exceed a factor of 2. To provide a safe therapeutic window larger than 2.5, pulse durations should exceed 10 ms for 330-μm beam diameter, and 20 ms for the 132-μm spot size.

It is important to keep in mind that coagulation of blood vessels requires more energy than other tissue does owing to cooling by the blood flow. For example, if a 100 ms exposure is applied in a 100 μm spot to occlude a blood vessel with flow velocity of 5 mm s^{-1}, the laser energy is effectively distributed over the 0.5-mm length of the blood vessel, which is five times

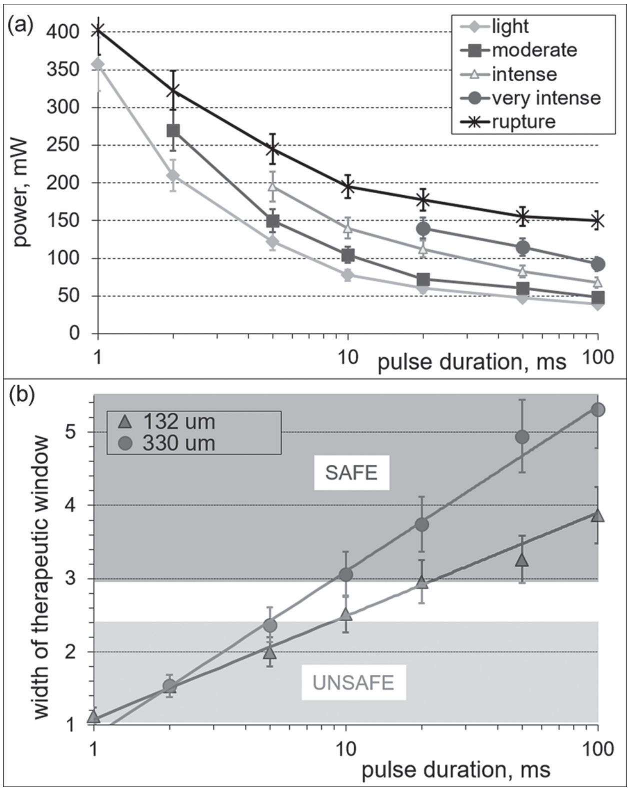


FIGURE 15.5 (a) Laser power required to create retinal lesions increases with decreasing pulse duration (measured with 132-μm spots on the retina). (b) The range of powers between photocoagulation and rupture (the therapeutic window) decreases with decreasing pulse duration, making visible photocoagulation unsafe with pulses shorter than 10 ms. (Figure adapted from Ref. [11].)

larger than the width of the laser spot. Thus, the temperature at the photocoagulation site is significantly lower than it would be in stationary tissue.

15.2.3.3 Pattern-Scanning Retinal Photocoagulation

For clinical efficacy of photocoagulation with smaller and lighter lesions to be maintained, a larger number of spots should be applied to provide the same total coagulated area [16]. Manual application of hundreds or even thousands of spots is fatiguing for both physician and patient, and maintaining constant spot density is rather difficult. The first attempts to automate photocoagulation involved rather complex equipment, including image-recognition software and eye-tracking [17]. Complexity of such systems prevented their commercial introduction and acceptance in clinical practice.

A semi-automatic pattern-scanning photocoagulator (PASCAL, Topcon Medical Laser Systems Inc.) was introduced by OptiMedica Corp. in 2005 [18]. It delivers patterns of laser spots in a rapid succession, activated by a single depression of a foot pedal. Patterns include square arrays with up to 5×5 spots, arcs with adjustable numbers of concentric rows, circular patterns for photocoagulation of small retinal holes, and rings and arcs with adjustable central exclusion zones for application around the fovea. To deliver the whole pattern within the eye fixation time, each exposure should be shorter than in conventional photocoagulation: 10–20 ms instead of 100–200 ms, traditionally applied with single spots.

Because lighter lesions produced with shorter pulses are smaller than more intense burns produced with longer exposures, applying more of them and at higher density is important to ensure the same total area is coagulated in PRP. For example, instead of 1000 moderate-grade lesions produced with 400-μm laser spots using 100-ms pulses, one needs to apply 1461 lesions of the same grade with 20-ms pulses, 1971 light-grade lesions with 20-ms pulses, or 3528 barely visible lesions with 20-ms exposures [11].

Even more automated laser delivery, guided by diagnostic imaging and stabilized using eye tracking, has been introduced in a Navilas system (OD-OS GmbH). This system includes retinal image acquisition, annotation of the images to create a detailed treatment plan, and then automated laser delivery to the retina according to the treatment plan. This system is particularly useful for image-guided targeting of the leaking microaneurysms [19].

15.2.3.4 Non-damaging Laser Therapy of the Macula

In the past, laser photocoagulation was used to treat extrafoveal choroidal neovascularization (CNV), which occurs in wet age-related macular degeneration (AMD). Intense photocoagulation destroys the invading vasculature but usually leaves a chorioretinal scar and associated blind spot (scotoma). Even though the treatment outside the centre of the macula is typically well tolerated by the patients, currently, a vast majority of physicians use anti-VEGF pharmacotherapy because it avoids scarring in the macula. Similarly, anti-VEGF injections are replacing the laser coagulation of the leaking microvascular abnormalities in branch retinal vein occlusion (BRVO).

Grid photocoagulation for treatment of the diabetic macular edema (DME) and central serous chorioretinopathy (CSCR) was shown to be efficient, but its mechanism of action was never understood. There is a growing body of clinical evidence that many macular diseases, such as CSCR, DME, and BRVO, can be successfully treated without tissue damage.

The non-damaging approach to retinal laser therapy was initially attempted using near-infrared diode laser (810 nm) with very long exposures (60 s) and a millimetre-wide spot on the retina [20]. This approach, termed transpupillary thermotherapy (TTT) [21], has been tested in the treatment of CNV in AMD [20,21]. Proponents of this approach have hypothesized that there is a selective damaging effect of heating on actively dividing cells in newly formed blood vessels owing to their higher susceptibility to thermal injury than non-dividing cells have in normal tissue. The estimated retinal temperature elevation with TTT at clinical settings (810-nm, 800-mW, 60-s, 3-mm spot size) is approximately 10°C [22]. The mechanism of treatment of CNV by TTT may occur through vascular thrombosis, apoptosis, or the thermal inhibition of angiogenesis [22]. The users of TTT encountered difficulties with reliable titration, resulting in frequent occurrences of significant retinal damage [23].

Later, a pulsed version of a similar laser with smaller spot size (125 *μ*m) was applied to non-damaging retinal therapy. The so-called micropulse laser delivers 100–300 ms long bursts composed of 0.1–0.3 ms pulses, repeated at a 500 Hz rate. The average power during the burst is set below the clinically detectable tissue damage by adjusting the peak power and pulse duty cycle (pulse duration divided by pulse period) in the range of 5%–15% [24].

Clinical trials have shown that micropulse treatment of DME delivered with high spot density is equally efficient as or superior to the standard modified Early Treatment Diabetic Retinopathy Study protocol [25]. Other clinical trials demonstrated that the micropulse-laser treatment reduced the sub-retinal fluid and improved visual acuity in patients with CSCR, compared to untreated control groups [26]. The procedure also exhibited equivalent clinical efficacy to conventional lasers in the treatment of macular edema secondary to BRVO but without the tissue damage [27]. However, the lack of a well-defined titration procedure is reflected in variable results of these studies. In addition, high-density coverage of the macula with small spots and long pulses requires a lengthy treatment.

Significant advantages of the non-damaging retinal phototherapy are the absence of scotomata and scarring, the ability to treat foveal areas, and improved preservation of colour vision and contrast sensitivity. The lack of chorioretinal damage permits high-density coverage, which greatly improves therapeutic outcomes compared to conventional sparse laser treatment protocols in the macula [25]. Nearly confluent laser applications could be safely delivered over the entire edematous areas if short-pulse treatment and pattern scanning are applied. This approach would also allow retreatment of the same areas, even in the fovea.

Dynamic range of the retinal response to non-damaging hyperthermia has been established by monitoring expression of the heat shock protein (HSP) HSP-70 in mice [28]. Based on these results, a titration protocol for adjustment of the laser power and duration was developed [29]. This protocol, called Endpoint Management (EpM), ties the sub-visible tissue effects to a visible titration. Laser power is first titrated to produce a barely visible lesion at 15 or 20 ms pulse duration, and this energy is set to be 100% on the EpM scale. Experiments with HSP expression following non-damaging retinal exposures in mice [28], as well as a computational analysis of the clinical laser settings [30], indicated that non-destructive thermal therapy corresponds to Arrhenius values within the range of approximately $0.1 < \Omega < 1$ (Figure 15.6). In this regime, the RPE cells survive the hyperthermia and respond to thermal stress by expression of HSPs. At higher settings ($\Omega > 1$), the RPE gets damaged, but below $\Omega \sim 100$, the lesions are ophthalmoscopically invisible, at least not immediately after the laser application. Immediately visible lesions produced at higher laser settings result in lethal damage to RPE and photoreceptors and have calculated values of $\Omega > 1000$ (Figure 15.6). By adjusting the pulse duration and power, the EpM algorithm maps logarithmic steps (10×) of Arrhenius integral values to linear steps (20%) in pulse energy, normalized to a titration dose [29], as shown in Figure 15.6.

With 200 µm spot size, the threshold of RPE damage was found to be 30% on the EpM scale [29]. Operating below the damage threshold allows the application of high-density patterns to increase clinical efficacy. Spot spacing of 0.25 spot diameter (50 µm) is typically applied. A treatment zone with a radius of 3 mm includes approximately 400 spots. This non-damaging protocol has been recently tested in the treatment of chronic CSCR using 30% energy settings. The trial demonstrated lack of tissue damage and excellent response to treatment (Figure 15.7) and to retreatments [31]. Additional clinical trials of this methodology in applications to DME, CSCR, macular telangiectasia, and dry AMD are in progress.

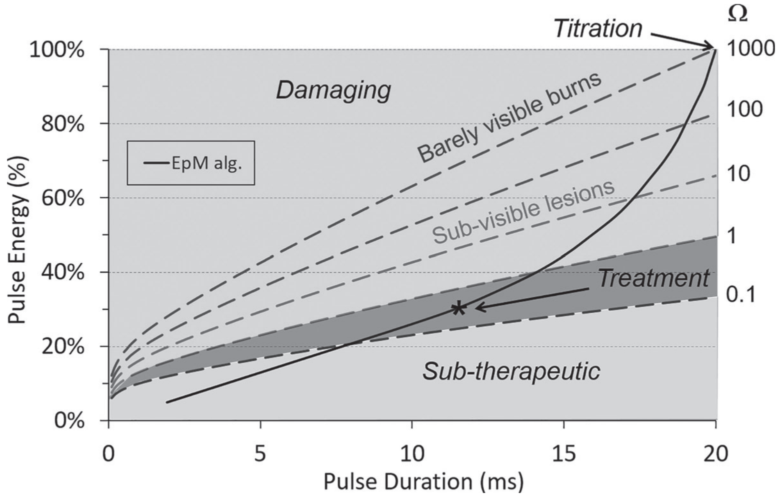


FIGURE 15.6 EpM protocol. Dashed lines correspond to lesions of various clinical grades, with the Arrhenius integral $\Omega = 1000$ corresponding to barely visible lesions, $\Omega = 1$ corresponding to RPE damage threshold, and $\Omega = 0.1$ corresponding to the lower end of the HSP expression range. When the laser power and pulse duration are adjusted, EpM (solid line) relates 10× steps in Ω to 20% steps in pulse energy. Damaging range corresponds to $\Omega > 1$. Non-damaging treatment range corresponds to $0.1 < \Omega < 1$. And sub-therapeutic range - to $\Omega < 0.1$. The 30% energy on the EpM scale in the middle of the non-damaging treatment range is shown by an asterisk. (Figure adapted from Ref. [11].)

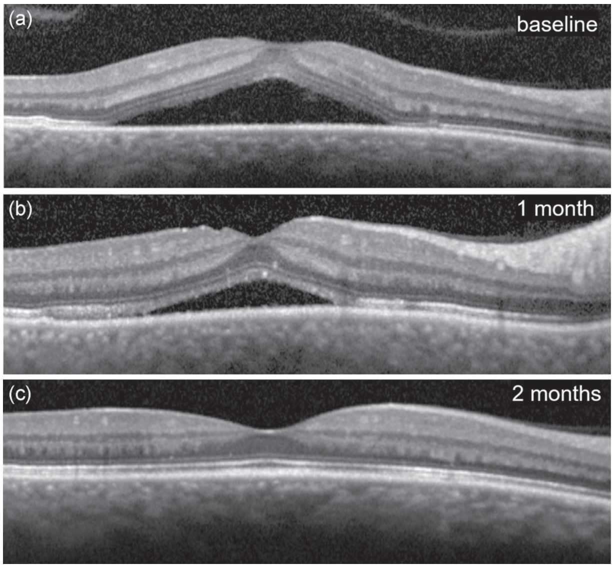


FIGURE 15.7 Resolution of sub-retinal fluid in a patient with chronic CSCR after NRT. Approximately, 400 spots have been applied at 30% EpM energy, and no tissue damage has been detected during the 12-month follow-up. (a) Visual acuity at baseline was 20/100, and it improved to (b) 20/40 at 1 month and (c) to 20/20 at two months. (Figure adapted from Ref. [11].)

15.2.3.5 Real-Time Monitoring of Tissue Temperature

Due to the strong variation in fundus pigmentation and transparency of the ocular tissues from patient to patient, and even between different areas in the same eye, the same laser settings may lead to very different results. Therefore, direct measurement of retinal temperature during photocoagulation would help for providing uniform outcomes, especially when the treatment is sub-visible.

A non-invasive method of measuring RPE temperature has been developed, based on detection of acoustic waves generated in melanosomes irradiated with nanosecond laser pulses [32]. An acoustic transducer built into a contact lens, which is attached to the treated eye during the procedure, detects the pressure waves generated due to thermoelastic expansion of melanosomes upon absorption of nanosecond laser pulses. The thermoelastic expansion coefficient of water varies with temperature by approximately 1% per 1°C [32], and this effect allows for monitoring the changes in temperature of the RPE cells by detecting the changes in amplitude of acoustic waves generated by the laser pulses of constant energy. The probing laser pulses are applied simultaneously with application of a therapeutic laser to detect temperature rise in tissue during the exposure. Precision of this method is on the order of 1°C, and this feedback is then used to adjust laser power or pulse duration to ensure predictable outcome of photocoagulation [33].

15.2.3.6 Laser Trabeculoplasty

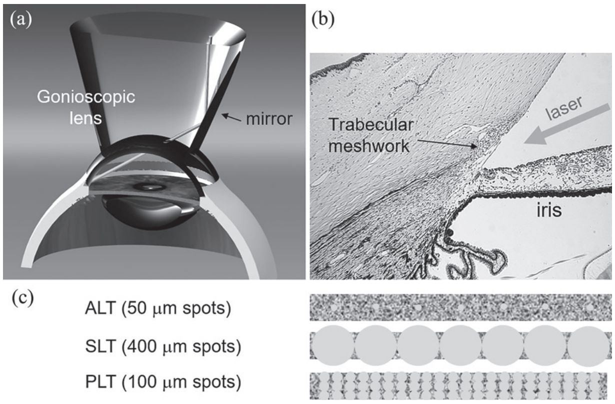


FIGURE 15.8 (a) Laser is guided onto the TM via gonioscopic lens for the treatment of the open-angle glaucoma. (b) Histology of the anterior angle of the human eye with illustration of the incoming laser beam directed onto the TM. (c) Diagram of the laser spots on TM: 50 μm spots of ALT, separated by 400 μm; adjacent to 400 μm spots of SLT; 3 rows of 100 μm spots of PLT, overlapping vertically by 25 μm, and separated laterally by 25 μm.

With the advent of the slit-lamp-based laser delivery system, the option of treating the iris and TM was an obvious application, and argon laser was first introduced for iridectomy [34] and trabeculoplasty (Figure 15.8) [35] in the 1970s. Both procedures were designed to create holes in the iris and in TM to improve aqueous outflow and reduce intraocular pressure (IOP) in cases of closed- and open-angle glaucoma, respectively. Safety and efficacy of the argon laser trabeculoplasty (ALT) in subjects with newly diagnosed primary open-angle glaucoma were demonstrated in 1995 [36]. With argon laser (514 nm) or more recently with the equivalent 532-nm Nd:YAG laser, 50 spots of 50 μm in diameter are applied to the 180° on TM with pulses of 100 ms in duration. IOP decreases by 25%–30%, and the effect lasts for several years, albeit it diminishes over time. Because laser burns in TM are much larger than the beam size and they scar over time, retreatment of the same areas is not possible.

Patterned laser trabeculoplasty (PLT) has recently been introduced for sub-visible or non-damaging treatment [37]. Its computer-guided patterns provide dense coverage of TM with 5 ms-long sub-visible exposures—a strategy similar to the non-damaging retinal laser therapy (NRT) described above. As with NRT, laser power is first titrated to a barely visible burn in the area of highest pigmentation (the inferior segment) using 10 ms pulses, and then pulse duration is decreased by half to reduce the energy below the visible damage threshold. Reduction in IOP following this treatment (~25%) was similar to the results of ALT, and owing to lack of tissue scarring, it allows periodic retreatments.

15.3 Tissue-Selective Therapy Using Photochemical Interactions

Photochemical interactions are based on non-thermal light-induced chemical reactions. Photochemical interactions used in PDT take place at very low power densities (typically <1 W cm^{-2}) and minute-long exposures to avoid heat accumulation. For PDT, special chromophores, called photosensitizers, are injected intravenously and allowed to accumulate in the target tissues. Upon excitation by laser radiation, the photosensitizer produces highly cytotoxic reactants—free radicals or singlet oxygen, which, in turn, produce irreversible oxidation of the nearby cell structures.

The concept of treating exudative AMD by selectively targeting vascular endothelial cells using a specific photosensitizer-carrier complex activated by near-infrared laser was adapted from tumour therapy [38,39]. Following successful initial results with phthalocyanine, liposomal benzoporphyrin derivative (verteporfin or Visudyne) was developed, which selectively attaches to the endothelium of new blood vessels [40]. This differential accumulation is the basis for selectivity of PDT against neovascularization, compared with the normal choroidal and retinal vasculature.

The far-red peak (688–691 nm) of the verteporfin absorption spectrum is typically utilized in clinical practice because of the lower retinal sensitivity and its superior penetration into the choroid. Closure of the abnormal blood vessels occurs within approximately 6–12 weeks in most patients [41]. Reperfusion is common, and multiple treatments are often required.

Since the recent advent of anti-VEGF pharmacotherapy, which has proven to be much more efficient in prevention of neovascularization, PDT has fallen out of favour, and its infrequent use in the United States is now limited to chronic CSCR [42].

More detailed description of PDT is provided in a separate chapter “Photodynamic Therapy”.

15.4 Photomechanical Interactions

Vapour bubbles produced when tissue temperature exceeds the vaporization threshold may rupture cells within a zone comparable to the bubble size. The actual temperature required for vaporization varies between 100°C and 305°C, depending on pulse duration and on the presence of the bubble nucleation sites [43]. For efficient heating, the energy should be delivered fast enough to avoid heat diffusion during the pulse, a condition called thermal confinement. In other words, laser pulse duration τ should be shorter than the time required for heat diffusion from the zone of laser absorption: $\tau < L^2/4\alpha$, where L is the light penetration depth ($L = 1/\mu_a$) and α is the thermal diffusivity of tissue. For example, with L = 1 μm, τ should not exceed 1.7 μs. Precise tissue ablation requires the use of laser wavelengths corresponding to a small optical penetration depth in tissue in order to confine the energy deposition to a small volume.

Much shorter pulses may produce stress confinement—a condition in which an acoustic wave generated by thermal expansion of the material cannot escape from the heated zone during the laser pulse. Since the velocity of sound in water is approximately $v = 1.5$ km s^{-1}, the stress confinement within the $L = 1$ μm is achieved if pulse duration does not exceed $\tau = L/v =$ 0.7 ns. Such conditions may promote the generation of powerful stress waves, which may result in the tissue disruption [43].

Vaporization of water in an overheated volume results in the formation of a short-living (microseconds) vapour bubble that expands, cools down, and collapses (cavitation) during the time determined by its maximum radius. These bubbles are not visible to the eye during surgery, but they can be easily visualized using fast flash photography.

Explosive vaporization produced by microsecond pulses of Er:YAG laser ($\lambda = 2.94$ μm, $L \approx 1$ μm in water) has been applied to the dissection of epiretinal membranes in front of the fibre probe [44]. Alternatively, overheating of liquid with nanosecond argon fluoride (ArF) excimer laser (λ = 193 nm) strongly absorbed in proteins ($L \approx 0.2$ μm in tissue) has been applied to the dissection of epiretinal and sub-retinal membranes [45]. Despite the early promise of these devices in clinical tests, both have failed to achieve acceptance in practice owing to the cost, fibre rigidity, and lack of coagulation capability.

15.4.1 Selective RPE Therapy

Light is strongly absorbed by melanosomes in the RPE ($\mu_a \approx 8000$ cm^{-1}). Application of microsecond laser pulses allows for confinement of the thermal and mechanical effects of this absorption within the RPE layer, thus sparing the photoreceptors and the inner retina [46]. Microsecond and nanosecond pulses can selectively damage RPE by the formation of small cavitation bubbles around melanosomes [6]. Microsecond exposures can also be conveniently produced by rapid scanning of a continuous laser beam. For example, scanning a beam of $d = 100$ μm in diameter with a speed of $v = 10$ m s^{-1} produces exposures of $t = d\ /\ v = 10$ μs in duration on the epithelial cells, sufficiently short for selective treatment of the RPE [47] (Figure 15.9a and b). Subsequent RPE proliferation and migration restore continuity of the RPE layer within a few days (Figure 15.9d and f).

Several clinical studies have demonstrated the efficacy of selective RPE therapy (SRT) in the treatment of DME, CSCR, and sub-foveal fluid after rhegmatogenous retinal detachment [48]. Despite its clinical promise, this technique has not been yet approved for clinical use. One of the difficulties with SRT is the lack of visible change in the retina, making the assessment of adequate laser dosimetry in every patient difficult. An acousto-optical system has been developed that may help assess the cavitation threshold energy in RPE [49]. Alternatively, light scattering by cavitation bubbles measured with a short flash could be used for this diagnostic.

15.4.2 Selective Laser Trabeculoplasty

Selective laser trabeculoplasty (SLT) was introduced in 1995 [50] by Mark Latina in the Wellman Labs at approximately the same time as SRT was developed in an adjacent room. As in SRT, the short pulses of light absorbed in pigmented cells cause explosive vaporization of melanosomes, leading to their destruction, while sparing the nearby non-pigmented tissue. SLT systems include a Q-switched, 532-nm laser that delivers 3 ns pulses in a 400 μm spot. In contrast to ALT, which results in scarring of the TM, SLT leaves the meshwork intact, with minimal damage to endothelial cells lining the meshwork beams [51]. With a 400 μm beam, 100 spots provide practically complete coverage of the TM (Figure 15.8c). As in SRT, epithelial cells migrate and proliferate to repopulate the TM, and it is hypothesized that in this process, the deposits clogging the TM are removed. SLT has been shown effective in the

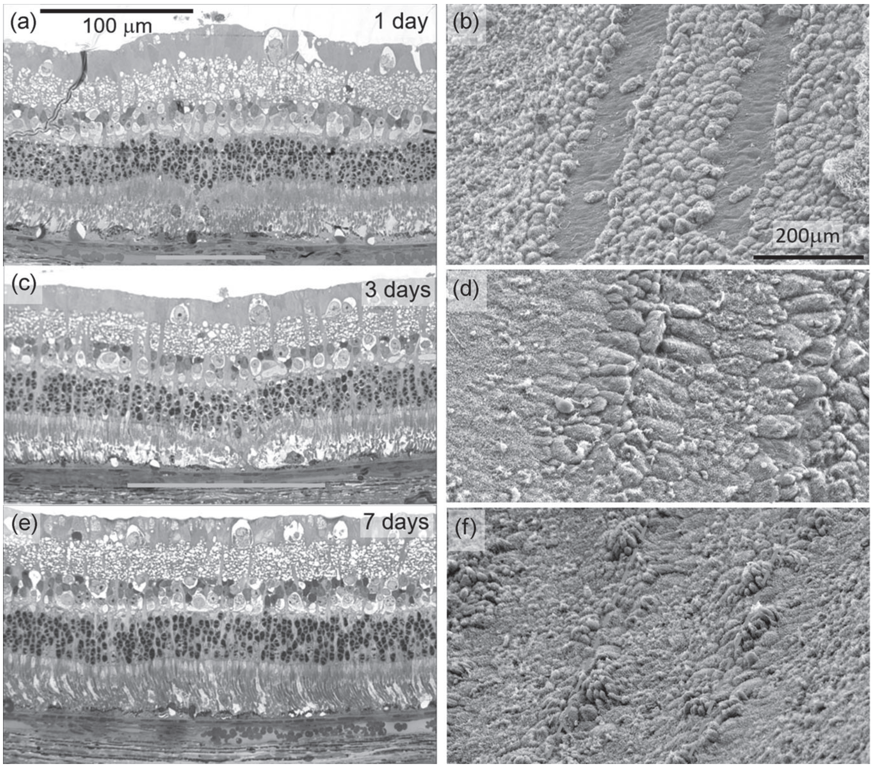


FIGURE 15.9 (a) RPE is selectively damaged by a rapidly scanning 532-nm laser. With a spot size of 100 μm and scanning velocity of 6.6 m s^{-1}, the exposure time is 15 μs. (b) Scanning electron microscopy (SEM) one day after the treatment reveals the laser scanning lines surrounded by the intact RPE. (c) Three days after treatment, RPE stretches into the lesion, but outer segments of photoreceptors are still abnormal. (d) SEM shows the damaged areas covered by the RPE migrated and stretched from the intact areas. (e) Seven days after treatment, RPE and outer segments are restored. (f) SEM shows complete, albeit uneven, coverage of RPE in the treated areas, including smaller, newly proliferated RPE cells. (Figure adapted from Ref. [11].)

treatment of open-angle glaucoma [51], thereby proving that permeability of TM to aqueous flow can be improved without destruction of its microstructure. SLT is easier to perform owing to its larger spot size and is better tolerated by patients because of reduced pulse energy. Similar to that of ALT, the IOP-lowering effect of SLT lasts for several years, but it also diminishes over time. However, unlike in ALT, lack of scarring in SLT allows retreatment.

15.4.3 Corneal Ablation for Refractive Surgery

The first ideas for reshaping the cornea to correct refractive errors were published by Lendeer Jans Lans from Holland in 1896, when he proposed using penetrating corneal cuts to correct astigmatism. In 1930, the Japanese ophthalmologist Tsutomu Sato demonstrated that radial cuts in the cornea could indeed correct up to six diopters. However, this procedure was not accepted in practice owing to the high rate of associated corneal degeneration. In 1974, Svyatoslav N. Fedorov (also spelled Fyodorov) in Moscow, Russia, began popularizing this procedure for the correction of myopia. The first attempts at corneal carving—a technique termed keratomileusis—started in 1963 in the Ignacio Barraquer clinic in Bogotá, Colombia. But the initial surgical manoeuvres were inherently imprecise, and modern refractive surgery dates to an accidental discovery in 1980.

Rangaswamy Srinivasan, a scientist in James Wynne's group at the Watson Research Center of IBM, was studying the application of a newly developed 193-nm ArF excimer laser to material processing. In October 1981, Srinivasan and Wynne put turkey cartilage left over from Thanksgiving dinner under the laser beam and observed a beautiful crater formed by ablation—much cleaner than any other laser they had tested. The process exhibited a previously unattainable level of exactitude—down to 0.2 μm. Owing to lack of thermal damage, the process was originally believed to be purely photochemical, whereby absorption of high-energy photons (E = 6.4 eV) led to decomposition of molecular bonds and material ejection because of the increased volume of the photofragments [52]. The process was termed photoablation (initially, ablative photodecomposition) to reflect this new mechanism, and it is described as such in many textbooks until today. However, further research revealed that the quantum yield of photochemical bond-breaking is insufficient to explain ablation thresholds and that vaporization of water in the cornea is actually responsible for material ejection [53]. Lack of detectable thermal damage was due to (i) very shallow penetration depth in tissue (0.2 μm in the cornea) and (ii) thermal confinement of the very short (10-ns) pulses [54,55].

Stephen Trokel and Francis L'Esperance of Columbia University first thought to use this laser as a replacement for a scalpel in radial keratotomy; however, a revolutionary approach of direct ablation of the corneal surface (termed photorefractive keratectomy, or PRK) has proven to be the most promising, and the first procedure was performed in 1983 [56].

Later, problems of slow and rather uncomfortable recovery of the epithelial layer were eliminated by first cutting a flap in the cornea, which is pulled back to expose the corneal bed for laser ablation—a procedure widely used today in clinical practice called laser-assisted in situ keratomileusis.

Another revolutionary step in refractive surgery was the development of the femtosecond laser for corneal flap cutting, pioneered by Tibor Juhasz's group in 1998 [57]. The mechanism of plasma-mediated interactions of ultra-fast lasers with transparent tissue is described in the next section. Unlike mechanical microkeratome, laser cutting enabled formation of vertical walls around the planar flap, and this process provided better positioning of the corneal flap back into the original location after ablation. This improved the consistency of refractive outcomes and led to wide acceptance of the femtosecond laser in refractive surgery. Later, ultra-fast lasers would also enable refractive surgical procedures based on intrastromal cutting without the excimer laser ablation: extraction of lenticules [58] and producing pockets for intrastromal rings [59]. The same laser systems have been applied to transplantation of the whole cornea and corneal endothelium [60], or the so-called endothelial keratoplasty.

A more detailed description of these applications is provided in a separate chapter "Refractive Surgery".

15.5 Transparent Tissue Surgery with Ultrashort-Pulse Lasers

Transparent tissue can be dissected utilizing a non-linear process of material ionization by a high-intensity laser beam. At extremely high irradiances (10^8–10^{11} W cm^{-2}), which can be achieved in a short-pulsed (nanosecond to femtosecond) tightly focused laser beam, transparent material can be ionized, and ions absorbing the laser light reach very high temperatures [43]. This mechanism, called dielectric breakdown, allows for highly localized energy deposition at the focal point of the laser beam, even inside a transparent medium. Threshold energy of the dielectric breakdown decreases with pulse duration and reaches its minimum around 100 fs [43]. Therefore, the most delicate surgical procedures are performed with lasers having pulse duration in the range of 100 fs.

15.5.1 Refractive Surgery

Dielectric breakdown with pulse durations in the range of 0.1–1 ps is applied to intrastromal dissection—formation of a corneal flap for refractive surgery [57], as well as in the cataract surgery [61].

In addition to corneal reshaping, femtosecond lasers can produce an opaque plane in the corneal stroma with a small aperture on the optical axis, which limits the pupil size in order to extend the focal depth in patients with presbyopia. In this application, laser cuts a planar ring-shaped pocket in the cornea, where a dye is injected to produce a dark tattoo. This novel procedure can replace implantation of the small-aperture inlay for the correction of presbyopia.

Another novel approach to refractive correction is based on adjustment of the corneal refractive index using molecular changes induced by multiphoton absorption of femtosecond laser in collagen [62]. Although the changes in refractive index are rather small (5–15×10^{-3}), grating patterns were shown to produce diffractive lenses with 1.5–2 diopters of power. Such patterns in cat corneas have been shown to be stable for at least 12 months.

15.5.2 Vitreoretinal Surgery

In addition to the anterior chamber of the eye, ultra-fast lasers have also been tested in the dissection of epiretinal membranes with a tightly focused beam [63]. Despite the very low energy requirements for this process (several microjoules with picosecond to femtosecond lasers), its applicability in the posterior pole is limited owing to the difficulty with axial discrimination between the epiretinal membranes and the retina located very close behind them. In addition, strong optical aberrations in the periphery of the posterior pole preclude tight focusing of the laser beam in these areas. However, further from the retina and from the posterior lens capsule, ultra-fast lasers could be applied to dissection of the vitreous floaters [64], especially if the pattern-scanning laser would be guided by three-dimensional imaging to ensure safe separation from the surrounding tissues.

15.5.3 Cataract Surgery

Cataract surgery is one of the most common surgical procedures in the United States, with approximately 2.5 million operations performed annually. Lasers have long played an important role in treatment of the posterior lens capsule opacification, called secondary cataract. Posterior capsule opacification occurs in approximately a third of the cases, and it was incised surgically for more than a century. Since 1974, the potential complications of bleeding and infection associated with such an open procedure have been avoided using the slit lamp-coupled nanosecond Nd:YAG laser (Figure 15.10) [65]. Newer picosecond and femtosecond lasers have a lower threshold energy of dielectric breakdown, with accordingly reduced

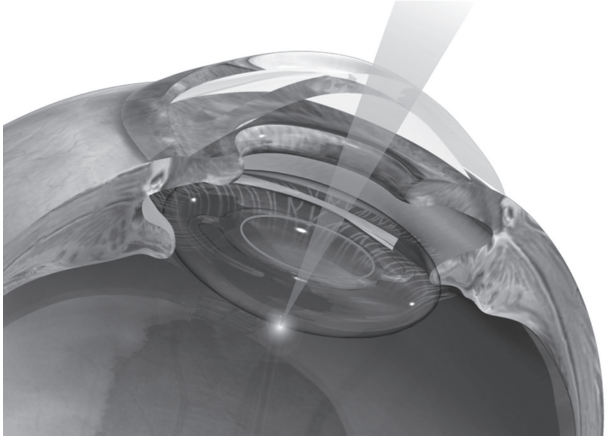


FIGURE 15.10 Dielectric breakdown at the focal point of a laser focused near the posterior capsule generates a transient cavitation bubble, which creates an opening in the opascified capsule on optical axis of the eye, thereby restoring its transparency.

size of the cavitation bubble and associated tissue damage and therefore allow more precise surgery.

Ultrashort-pulse lasers have been applied to softening of the aging crystalline lens in an attempt to restore accommodation. Initial attempts of generating intra-lenticular incisions to soften the lens tissue and re-establish its flexibility in presbyopia were performed with nanosecond laser by Myers and Krueger [66]. Threshold energy and associated generation of the residual gas bubbles were later reduced with application of femtosecond pulses. Recent investigations have shown softening of the crystalline lens by femtosecond laser [67], and with properly selected laser settings, this process has not induced cataract formation. Clinical trials are now in progress to assess the extent of improvement in accommodation in human patients undergoing laser lentotomy.

Recently, a new laser-based technique has been developed for anterior capsulotomy (a circular cut in the anterior lens capsule to create an opening for removal of the crystalline lens and implantation of the intraocular lens) and lens segmentation [61]. A scanning femtosecond laser can produce fine-cutting patterns in tissue, with their placement defined by integrated optical coherence tomography (OCT) imaging (Figure 15.11). Laser incisions in the anterior capsule are much more precise in size and shape than manual capsulorhexis (manual tearing of the anterior lens capsule in a circular pattern to create an opening for removal of the crystalline lens and implantation of the intraocular lens), as illustrated in Figure 15.12a,d and e, ensuring perfect centration of the intraocular lens and accurate overlap of the capsular bag with the implant [68]. Segmentation and softening of the lens (Figure 15.12b) simplify its subsequent ultrasonic emulsification, thereby reducing the collateral damage to corneal endothelium during lens removal [61]. Femtosecond lasers have also been applied for construction of multi-planar self-sealing cataract incisions (Figure 15.12c) to improve safety of the procedure and prevent infections, and for exact placement of limbal relaxing incisions to reduce residual astigmatism [61]. Image-guided laser surgery eliminates the uncertainty of the surgical skills, tissue visibility, and variability of its elasticity and strength in various patients, making perfectly sized and centred cuts in every case. However, the large size of these laser systems precluded their integration into the surgical microscopes, and therefore, laser dissection is performed as a separate procedure prior to cataract surgery. This additional step in cataract surgery slows the process and adds cost, thereby limiting the use of this technology in clinical practice. Smaller and less expensive lasers integrated into the surgical microscope would make this technology much more appealing and practical.

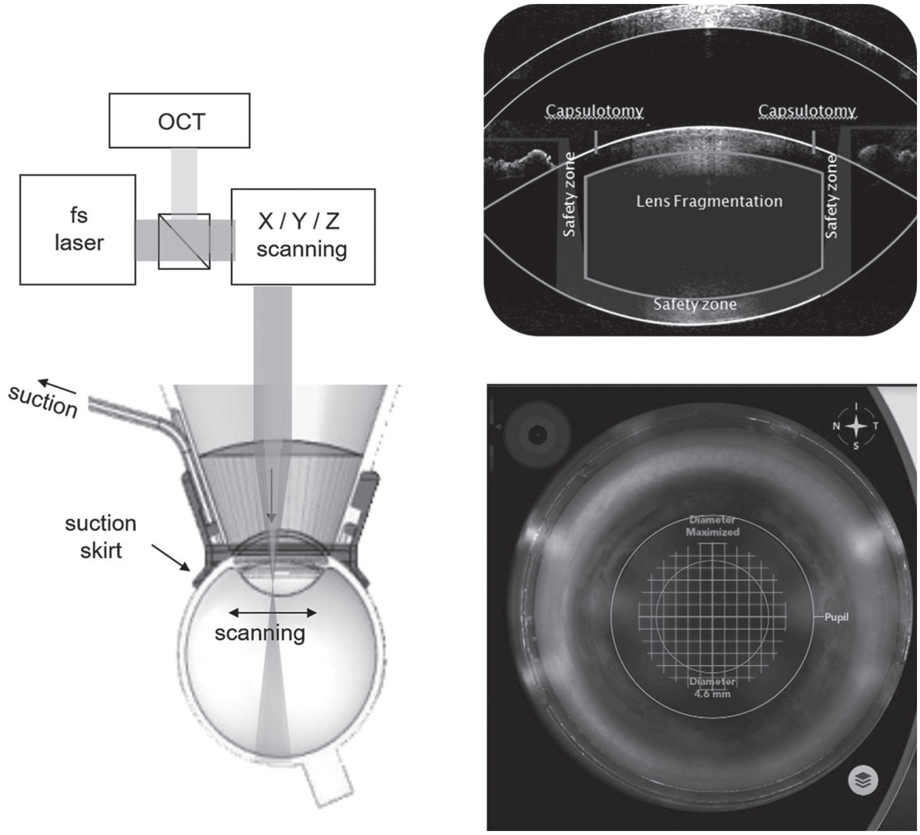


FIGURE 15.11 OCT imaging system integrated with a femtosecond laser for cataract surgery. Cutting patterns in the crystalline lens and in the anterior capsule are placed based on the 3-D image acquired by OCT.

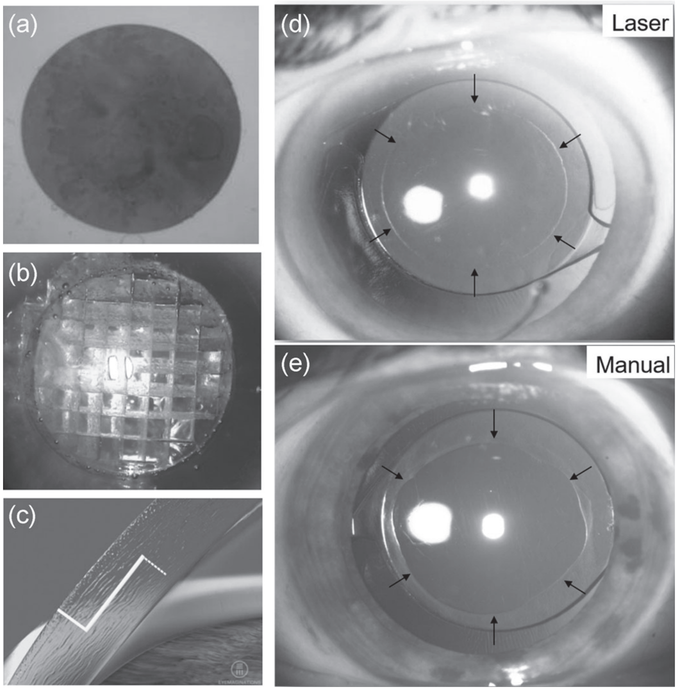


FIGURE 15.12 (a) A disk of the anterior capsule cut by laser during cataract surgery. (b) Crystalline lens segmented by femtosecond laser to ease its extraction during cataract surgery. (c) Multi-planar cut in the cornea produces the self-sealing incision, which reduces the fluid outflow during operation and prevents infections after cataract surgery. (d) An eye of a patient after the laser-assisted cataract surgery. Perfectly round edges of the laser capsulotomy are indicated by the arrows. (e) Patient's eye after conventional manual surgery, with very uneven capsulotomy pointed to by the arrows. Accurately shaped and centred capsulotomy helps ensure centration of the intraocular lens and perfect overlap of the capsule edges with the implant to prevent cellular proliferation into the posterior capsule, which leads to its opacification (secondary cataract). (Figure adapted from Ref. [11].)

15.6 Summary and Future Directions

Modern ophthalmic lasers enable unprecedented precision and selectivity in many therapeutic and surgical procedures and have become indispensable for ophthalmic care. Progress in understanding of the laser–tissue interactions and the tissue response pathways leads to more selective and less-damaging treatments, such as (i) minimally traumatic photocoagulation which utilizes retinal plasticity to minimize scarring; (ii) selective treatment of RPE (SRT) and trabecular epithelium (SLT) based on intracellular explosion of melanosomes using microsecond and nanosecond pulses; and (iii) non-damaging laser therapy of the macula using transient hyperthermia below the damage threshold. Advancements in laser and imaging technologies enable ultraprecise and fully automated surgery and therapy, such as refractive and cataract surgeries based on ultra-fast lasers, or image-guided photocoagulation with eye-tracking.

Variations of the retinal pigmentation as well as transparency of the ocular tissues and laser focusing affect the resulting temperature rise during treatment. Real-time dosimetry of the laser exposure on the retina or TM will ensure delivery of the desired dose in each spot, which is especially important for the non-damaging treatment. As technology advances, manual treatments will be replaced with more automated pre-planned image-guided retinal laser therapy. Novel retinal therapies, such as NRT and SRT, will have to be validated in treatment of various macular disorders, including CSCR, MacTel, DME, and AMD.

Large size and high cost of the ultra-fast lasers preclude their penetration into clinical practice. Smaller and less expensive ultra-fast lasers integrated with the surgical microscope in the operating room will allow broader use of this technology in cataract and refractive surgery. Formation of transscleral channels with ultra-fast near-infrared laser to allow controlled outflow of fluid from the eye for reduction of the IOP may replace more invasive manual glaucoma surgery. Currently, there is no solution for the vitreous floaters—clumps of the vitreous gel, which may obscure parts of the visual field. Image-guided ablation of the vitreous floaters using ultra-fast lasers may solve this problem.

Disclosure Statement

Author is an inventor of the Pattern Scanning Laser Photocoagulator (PASCAL). The patent is licensed by Stanford University to Topcon Medical Laser Systems (TMLS). Author is also a consultant to TMLS. Author is an inventor of the OCT-guided femtosecond laser system for cataract surgery (Catalys), which is currently manufactured by Johnson & Johnson. Author is also a consultant to Johnson & Johnson.

REFERENCES

1. Zaret, M.M., et al., Ocular lesions produced by an optical maser (laser). *Science*, 1961. **134**(348): pp. 1525–6.
2. L'Esperance, F.A., Jr., The treatment of ophthalmic vascular disease by argon laser photocoagulation. *Trans Am Acad Ophthalmol Otolaryngol*, 1969. **73**(6): pp. 1077–96, (Discussion by HC Zweng, pp 1095–1096).
3. Thompson, K.P., Q.S. Ren, and J.-M. Parel, Chapter 8: Therapeutic and Diagnostic Application of Lasers in Ophthalmology, in *Lasers in Medicine*, R.W. Waynant, Editor. 2002, CRC Press: Boca Raton, FL.
4. Palanker, D. and M.S. Blumenkranz, Chapter 39- Retinal Laser Therapy: Biophysical Basis and Applications A2- Ryan, Stephen J, in *Retina* (Fifth Edition), S.R. Sadda, et al., Editors. 2013, W.B. Saunders: London. pp. 746–60.
5. Sramek, C., et al., Dynamics of retinal photocoagulation and rupture. *J Biomed Opt*, 2009. **14**(3): pp. 034007–1–13.
6. Schuele, G., et al., RPE damage thresholds and mechanisms for laser exposure in the microsecond-to-millisecond time regimen. *Invest Ophthalmol Vis Sci*, 2005. **46**(2): pp. 714–9.
7. Sramek, C.K., et al., Therapeutic Window of Retinal Photocoagulation With Green (532-nm) and Yellow (577-nm) Lasers. *Ophthalmic Surg Lasers Imaging*, 2012: pp. 1–7.
8. Beetham, W.P., et al., Ruby laser photocoagulation of early diabetic neovascular retinopathy: preliminary report of a long-term controlled study. *Arch Ophthalmol*, 1970. **83**(3): pp. 261–72.
9. DRS_Study_Group, Photocoagulation Treatment for Proliferative Diabetic Retinopathy. Clinicial Application of DRS findings, DRS Report Number 8. *Ophthalmology*, 1981. **88**: pp. 583–600.
10. ETDRS_Study_Group, Early Photocoagulation for Diabetic Retinopathy. ETDRS Report Number 9. *Ophthalmology*, 1991. **98**: pp. 766–85.
11. Palanker, D., Evolution of concepts and technologies in ophthalmic laser therapy. *Annu Rev Vis Sci*, 2016. **2**: pp. 295–319.
12. Maeshima, K., et al., Progressive enlargement of scattered photocoagulation scars in diabetic retinopathy. *Retina*, 2004. **24**(4): pp. 507–11.
13. Sher, A., et al., Restoration of retinal structure and function after selective photocoagulation. *J Neurosci*, 2013. **33**(16): pp. 6800–8.
14. Muqit, M.M., et al., Pain responses of Pascal 20 ms multi-spot and 100 ms single-spot panretinal photocoagulation: Manchester Pascal Study, MAPASS report 2. *Br J Ophthalmol*, 2010. **94**(11): pp. 1493–8.
15. Jain, A., et al., Effect of pulse duration on size and character of the lesion in retinal photocoagulation. *Arch Ophthalmol*, 2008. **126**(1): pp. 78–85.
16. Palanker, D., et al., The impact of pulse duration and burn grade on size of retinal photocoagulation lesion: implications for pattern density. *Retina*, 2011. **31**(8): pp. 1664–9.
17. Wright, C.H.G., et al., Initial in vivo results of a hybrid retinal photocoagulation system. *J Biomed Opt*, 2000. **5**(1): pp. 56–61.
18. Blumenkranz, M.S., et al., Semiautomated patterned scanning laser for retinal photocoagulation. *Retina*, 2006. **26**(3): pp. 370–6.
19. Kozak, I., et al., Clinical evaluation and treatment accuracy in diabetic macular edema using navigated laser photocoagulator NAVILAS. *Ophthalmology*, 2011. **118**(6): pp. 1119–24.
20. Reichel, E., et al., Transpupillary thermotherapy of occult subfoveal choroidal neovascularization in patients with age-related macular degeneration. *Ophthalmology*, 1999. **106**(10): pp. 1908–14.
21. Newsom, R.S.B., et al., Transpupillary thermotherapy (TTT) for the treatment of choroidal neovascularisation. *Br J Ophthalmol*, 2001. **85**(2): pp. 173–8.
22. Mainster, M.A. and E. Reichel, Transpupillary thermotherapy for age-related macular degeneration: long-pulse photocoagulation, apoptosis, and heat shock proteins. *Ophthalmic Surg Lasers*, 2000. **31**(5): pp. 359–73.
23. Benner, J.D., R.M. Ahuja, and J.W. Butler, Macular infarction after transpupillary thermotherapy for subfoveal choroidal neovascularization in age-related macular degeneration. *Am J Ophthalmol*, 2002. **134**(5): pp. 765–8.
24. Dorin, G., Evolution of retinal laser therapy: minimum intensity photocoagulation (MIP). Can the laser heal the retina without harming it? *Semin Ophthalmol*, 2004. **19**: pp. 62–8.
25. Lavinsky, D., et al., Randomized clinical trial evaluating mETDRS versus normal or high-density micropulse photocoagulation for diabetic macular edema. *Invest Ophthalmol Vis Sci*, 2011. **52**(7): pp. 4314–23.
26. Roisman, L., et al., Micropulse diode laser treatment for chronic central serous chorioretinopathy: a randomized pilot trial. *Ophthalmic Surg Lasers Imaging Retina*, 2013. **44**(5): pp. 465–70.
27. Parodi, M.B., et al., Subthreshold grid laser treatment of macular edema secondary to branch retinal vein occlusion with micropulse infrared (810 nanometer) diode laser. *Ophthalmology*, 2006. **113**(12): pp. 2237–42.
28. Sramek, C., et al., Non-damaging retinal phototherapy: dynamic range of heat shock protein expression. *Invest Ophthalmol Vis Sci*, 2011. **52**(3): pp. 1780–7.
29. Lavinsky, D., et al., Subvisible retinal laser therapy: titration algorithm and tissue response. *Retina*, 2014. **34**(1): pp. 87–97.
30. Luttrull, J.K., et al., Long-term safety, high-resolution imaging, and tissue temperature modeling of sub-visible diode micropulse photocoagulation for retinovascular macular edema. *Retina*, 2011. **in print**.
31. Lavinsky, D., et al., Nondamaging retinal laser therapy: rationale and applications to the macula. *Invest Ophthalmol Vis Sci*, 2016. **57**(6): pp. 2488–500.
32. Framme, C., et al., Influence of pulse duration and pulse number in selective RPE laser treatment. *Lasers Surg Med*, 2004. **34**(3): pp. 206–15.
33. Koinzer, S., et al., Temperature-controlled retinal photocoagulation reliably generates uniform subvisible, mild, or moderate lesions. *Transl Vis Sci Technol*, 2015. **4**(5): p. 9.
34. Beckman, H. and H.S. Sugar, Laser iridectomy therapy of glaucoma. *Arch Ophthalmol*, 1973. **90**(6): pp. 453–5.

35. Teichmann, I., K.D. Teichmann, and P.U. Fechner, Glaucoma operation with the argon laser. *Eye Ear Nose Throat Mon*, 1976. **55**(6): pp. 58–62.
36. Glaucoma Laser Trial Research Group, The Glaucoma Laser Trial (GLT) and glaucoma laser trial follow-up study: 7. Results. *Am J Ophthalmol*, 1995. **120**(6): pp. 718–31.
37. Turati, M., et al., Patterned laser trabeculoplasty. *Ophthalmic Surg Lasers Imaging*, 2010. **41**(5): pp. 538–45.
38. Dougherty, T.J. and T.S. Mang, Characterization of intratumoral porphyrin following injection of hematoporphyrin derivative or its purified component. *Photochem Photobiol*, 1987. **46**(1): pp. 67–70.
39. Schmidt, U., R. Birngruber, and T. Hasan, [Selective occlusion of ocular neovascularization by photodynamic therapy]. *Ophthalmologe*, 1992. **89**(5): pp. 391–4.
40. Miller, J.W., et al., Photodynamic therapy of experimental choroidal neovascularization using lipoprotein-delivered benzoporphyrin. *Arch Ophthalmol*, 1995. **113**(6): pp. 810–8.
41. Schmidt-Erfurth, U., et al., Vascular targeting in photodynamic occlusion of subretinal vessels. *Ophthalmology*, 1994. **101**(12): pp. 1953–61.
42. Oh, B.L. and H.G. Yu, Choroidal thickness after full-fluence and half-fluence photodynamic therapy in chronic central serous chorioretinopathy. *Retina*, 2015. 35(8): pp. 1555–60.
43. Vogel, A. and V. Venugopalan, Mechanisms of pulsed laser ablation of biological tissues. *Chem Rev*, 2003. **103**(2): pp. 577–644.
44. D'Amico, D.J., et al., Multicenter clinical experience using an erbium:YAG laser for vitreoretinal surgery. *Ophthalmology*, 1996. **103**(10): pp. 1575–85.
45. Hemo, I., et al., Vitreoretinal surgery assisted by the 193-nm excimer laser. *Invest Ophthalmol Vis Sci*, 1997. **38**(9): pp. 1825–9.
46. Roider, J., et al., Microphotocoagulation: selective effects of repetitive short laser pulses. *Proc Natl Acad Sci USA*, 1993. **90**(18): pp. 8643–7.
47. Paulus, Y.M., et al., Selective retinal therapy with a continuous line scanning laser. *SPIE Proc*, 2010. **7550** (Ophthalmic Technologies XX): p. 75500W.
48. Brinkmann, R., J. Roider, and R. Birngruber, Selective retina therapy (SRT): a review on methods, techniques, preclinical and first clinical results. *Bull Soc Belge Ophtalmol*, **2006**(302): pp. 51–69.
49. Brinkmann, R., et al., Determination of absolute fundus temperatures during retinal laser photocoagulation and selective RPE treatment. *Invest Ophthalmol Visual Sci*, 2001. **42**(4): pp. S696–S696.
50. Latina, M.A. and C. Park, Selective targeting of trabecular meshwork cells: in vitro studies of pulsed and CW laser interactions. *Exp Eye Res*, 1995. **60**(4): pp. 359–71.
51. Latina, M.A., et al., Q-switched 532-nm Nd:YAG laser trabeculoplasty (selective laser trabeculoplasty): a multicenter, pilot, clinical study. *Ophthalmology*, 1998. **105**(11): pp. 2082–8; discussion 2089–90.
52. Garrison, B.J. and R. Srinivasan, Laser ablation of organic polymers - microscopic models for photochemical and thermal-processes. *J Appl Phys*, 1985. **57**(8): pp. 2909–14.
53. Kitai, M.S., et al., The physics of Uv laser cornea ablation. *IEEE J Quan Electron*, 1991. **27**(2): pp. 302–7.
54. Isner, J. and R. Clarke, The paradox of thermal ablation without thermal injury. *Lasers Med Sci*, 1987. **2**(3): pp. 165–73.
55. Telfair, W.B., et al., Histological comparison of corneal ablation with Er: YAG laser, Nd: YAG optical parametric oscillator, and excimer laser. *J Refract Surg*, 2000. **16**(1): pp. 40–50.
56. Trokel, S.L., R. Srinivasan, and B. Braren, Excimer laser-surgery of the cornea. *Am J Ophthalmol*, 1983. **96**(6): pp. 710–5.
57. Kurtz, R.M., et al., Lamellar refractive surgery with scanned intrastromal picosecond and femtosecond laser pulses in animal eyes. *J Refract Surg*, 1998. **14**(5): pp. 541–8.
58. Heisterkamp, A., et al., Photodisruption with ultrashort laser pulses for intrastromal refractive surgery. *Laser Phys*, 2003. **13**(5): pp. 743–8.
59. Ertan, A. and M. Bahadir, Intrastromal ring segment insertion using a femtosecond laser to correct pellucid marginal corneal degeneration. *J Cataract Refract Surg*, 2006. **32**(10): pp. 1710–6.
60. Jonas, J.B., Corneal endothelial transplantation using femtosecond laser technology. *Eye*, 2004. **18**(6): pp. 657–8.
61. Palanker, D.V., et al., Femtosecond laser-assisted cataract surgery with integrated optical coherence tomography. *Sci Transl Med*, 2010. **2**(58): p. 58ra85.
62. Ding, L., et al., Intratissue refractive index shaping (IRIS) of the cornea and lens using a low-pulse-energy femtosecond laser oscillator. *Invest Ophthalmol Vis Sci*, 2008. **49**(12): pp. 5332–9.
63. Cohen, B.Z., K.J. Wald, and K. Toyama, Neodymium:YLF picosecond laser segmentation for retinal traction associated with proliferative diabetic retinopathy. *Am J Ophthalmol*, 1997. **123**(4): pp. 515–23.
64. Gupta, R., et al., Nd:YAG laser photocystotomy of a free-floating pigmented anterior vitreous cyst. *Ophthalmic Surg Lasers Imaging*, 2003. **34**(3): pp. 203–5.
65. Aron-Rosa, D., et al., Use of the neodymium-YAG laser to open the posterior capsule after lens implant surgery: a preliminary report. *J Am Intraocul Implant Soc*, 1980. **6**(4): pp. 352–4.
66. Krueger, R.R., et al., Experimental increase in accommodative potential after neodymium: yttrium-aluminum-garnet laser photodisruption of paired cadaver lenses. *Ophthalmology*, 2001. **108**(11): pp. 2122–9.
67. Hahn, J., et al., Measurement of ex vivo porcine lens shape during simulated accommodation, before and after fs-laser treatment. *Invest Ophthalmol Vis Sci*, 2015. **56**(9): pp. 5332–43.
68. Friedman, N.J., et al., Femtosecond laser capsulotomy. *J Cataract Refract Surg*, 2011. **37**(7): pp. 1189–98.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

16

Therapeutic Application: Refractive Surgery

Leonardo Mastropasqua

CONTENTS

16.1 Introduction 243
16.2 Excimer Laser-Based Procedures 243
16.3 Excimer Laser-Based Procedures Complications 244
16.4 FSL-Based Procedures 245
16.5 ReLEx Procedures 246
16.6 Newer Application of Small-Incision Lenticule Extraction 246
References 247

16.1 Introduction

Refractive error (ametropia) is a worldwide very common condition resulting in decreased visual acuity. Eyeglasses, contact lenses, and surgery represented possible treatments and solutions for ametropia correction. In particular, laser eye surgery to correct refractive errors in the eye (myopia, hyperopia, astigmatism) has become increasingly diffuse and replaces other methods used to correct refractive errors.

Several laser refractive surgical procedures have been used to reshaping the cornea and correct myopia, hyperopia, astigmatism, and presbyopia.

Excimer laser introduction represented a very remarkable revolution in correction of these refractive errors, combined with other advanced ophthalmological instruments for customizing procedure and minimizing complications.

Photorefractive keratectomy (PRK) was developed over 20 years ago, and subsequently, wavefront-guided laser in situ keratomileusis (LASIK) has become the most commonly performed keratorefractive procedure.

Despite these advances, limitations in excimer laser ablation and subsequent complications persist. Femtosecond laser (FSL) technology has revolutionized corneal refractive surgery and gained widespread acceptance due to the versatility, precision, reproducibility, and safety of the instruments; first and major application of FSL in corneal refractive surgery is represented by flap creation for LASIK procedure. Since 2010, advances in FSL technologies allowed to obtain “all in one” refractive surgeries based on refractive corneal lenticule sculpture and extraction: femtosecond lenticule extraction (Flex) and SMILE procedures.

In this chapter, history of laser refractive surgery is discussed together with the presentation of state-of-the-art of “new refractive surgery era”.

16.2 Excimer Laser-Based Procedures

The different types of interactions of laser-produced light with tissue depend on the wavelength and this also determines the colour of that light.

The lasers used in corneal surgery produce infrared or ultraviolet radiations.

In particular, excimer (excited dimer) lasers belong to the latter group, producing wavelengths between 150 and 300 nm.

Dimers are atomic molecules that, in the excited state, are formed from a noble gas and a halogen, and which, when returning to the ground state, split and emit radiation in the ultraviolet field, thus liberating the energy previously used to form the dimer.

In corneal refractive surgery different dimers are obtained from the mixture of inert gas and halogen and these will produce different laser emissions.

To perform corneal surgery, the wavelength chosen is the one with a photonic energy which is able to break the interatomic links with the least possible damage to the tissues and is produced by the dimer argon fluoride (ArF) that emits at 193 nm.

The photonic energy of this radiation is able to split carbon links in the corneal tissues, where it is totally absorbed without any transmission phenomenon.

To obtain photoablation, the energy must be applied with high density for an extremely short time. Excimer lasers work in the pulsed mode; in fact, the half-life of the dimer is short enough to allow equally quick laser emission. With this characteristic the operative parameters are as follows: the energetic fluency is 160–200 MJ cm^{-3}, with a quantity of ablated stroma of 0.25 μm for the impulse.

The exposure time for each spot is about 9–14 ns; the emission frequency of the pulses is between 5 and 20 Hz (it is not

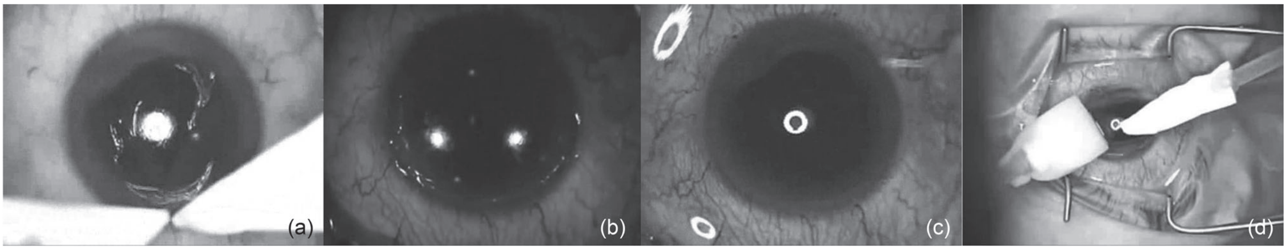


FIGURE 16.1 (a–d) PRK procedure: epithelial removal and excimer laser ablation.

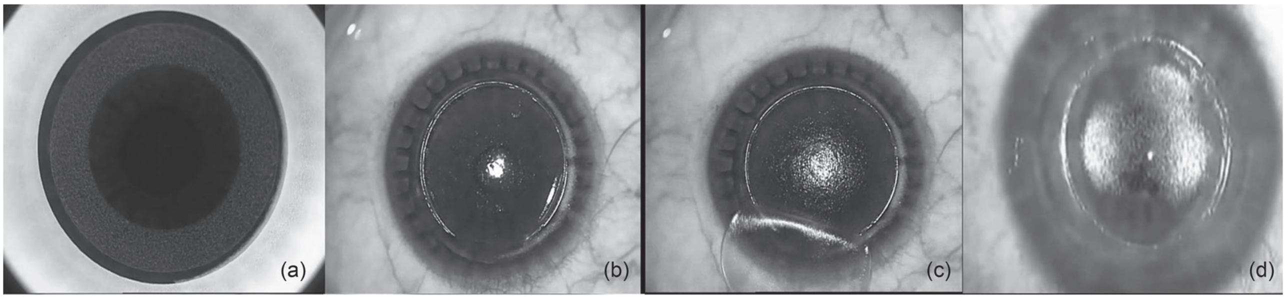


FIGURE 16.2 (a–d) LASIK procedure: flap creation and excimer laser ablation.

possible to use higher frequency because overheating could arise and photoablated particles need time to leave the exposed zone). The optic arm offers a constant nitrogen flow to control the air humidity and a system of mirrors, lenses, and prisms to prevent light radiation from dispersion [1].

During years, a wide range of surgical techniques that modify eye refractive error by removing corneal tissue and reshaping the cornea with excimer laser have been conceived. These procedures can be broadly divided into 2 categories: corneal surface ablation surgery and corneal stromal ablation surgery (involving the creation of corneal flap). Surface ablation procedures include PRK, transepithelial photorefractive keratectomy, laser epithelial keratomileusis, and epipolis laser in situ keratomileusis.

On the other side, corneal stromal ablation procedures include LASIK with the flap created with either a mechanical microkeratome or femtosecond-based microkeratome (FS-LASIK) [2–4].

In this chapter only PRK and LASIK are currently considered as more widespread excimer laser-based procedures.

During PRK corneal epithelium is removed, a large epithelial defect ensues, and after photoblation, healing occurs by migration and division of surrounding epithelium. The epithelium might be removed in several ways, including excimer laser destruction, mechanical debridement with a surgical blade, abrasion with a brush, or use of alcohol to loosen the epithelium [2] (Figure 16.1).

LASIK is lamellar laser refractive surgery in which the excimer laser ablation is done under a partial-thickness lamellar corneal flap. Until recently, the lamellar flap could only be made with a microkeratome. The microkeratome uses an oscillating blade to cut the flap after immobilization of the cornea with a suction ring that could cut to depths of 100–200 µm with either superior or nasal hinges. Recently FSL introduction in corneal surgery allows to create thinner flap at a desired corneal depth; therefore, FS-LASIK procedure is more accurate with reduction of flap-related complications since FSL flap creation is less dependent on corneal curvature and might be more reliable in cases of steep or flat corneas, which can be at greater risk for irregular corneal cuts [2,5] (Figure 16.2).

In both techniques, concentric portions of corneal stroma are ablated. The currently used laser is the flying-spot laser, which uses a smaller spot size than the previously used broad-beam lasers to correct hyperopia, myopia, and astigmatism. The laser creates very small (2 mm or less) round pulses of energy. Since the pulses are small, the laser energy can be very consistent and reproducible. In the treatment of myopias and myopic astigmatism, the laser is programmed to assure a higher exposure of the central corneal surface and, consequently, an increase in the corneal bending radius, a flattening of the corneal curvature, and a reduction in the refractive power of the external corneal surface [1].

In the treatment of hypermetropia and hypermetropic astigmatism, the laser preserves central zones while the peripheral surface is modelled (steepening of the corneal curvature). The frequency of application with flying-spot lasers is higher than with broad-beam lasers. For this reason, treatments using flying-spot lasers generally take much longer to complete. To maintain the position for a much longer time, flying-spot lasers are provided with eye-tracking devices [1].

16.3 Excimer Laser-Based Procedures Complications

Surface ablation procedures' complications may be classified as follows:

Infections: Infections following PRK are rare but potentially sight-threatening.

In literature bacterial infections were described with good results after adequate therapy.

Introduction of antibiotic prophylaxis should exert adequate cover against Gram-positive organisms [6,7].

Viral keratitis has also been described, probably related to a reactivation of latent HSV keratitis [8].

Ectasia: Corneal ectasia is a rare but well-described complication of LASIK, indicating an altered biomechanical strength, but has rarely been reported after PRK except in cases with features suggestive of ectasia or a family history of keratoconus and cases of multiple treatments and enhancements [9].

Haze: The most common complication after PRK treatment is represented by haze, an alteration in the corneal transmittance, visible by biomicroscopy as an opalescence.

Histologically, immediately after PRK treatment, a flat rather uniform area on the stromal surface is observed, with a wavy aspect of the order of some micrometres and without prominent irregularities. The exposed area is subsequently covered by a multi-stratified corneal epithelium where the basal epithelial layer adheres directly to the corneal stroma through a new basal membrane. After one month, it is possible to observe a superficial stromal reaction consisting in the deposition of disorganized neocollagen and in vacuolating phenomena of the epithelial–stromal junction [1].

From the clinical point of view, a superficial stromal reaction causes haze. It can last a varying time (3–6 months) and is not constant and equal in all cases: some of them present a greater reaction with a greater haze persistence whilst, in other cases, the corneal opalescence is extremely small and vanishes very quickly [1].

Many authors [10–12] have subjectively classified haze as follows: grade 0, completely transparent cornea; grade 0.5, scarcely perceptible haze; grade 1, minimal density haze at direct and reflected illumination; grade 2, light haze easily visible with focal direct slit lamp illumination; grade 3, moderate haze, partially obscuring irideal details; grade 4, marked haze, totally obscuring irideal details.

After normal scarring process, haze does not influence visual acuity but, at most, the contrast sensibility and only for the first months after the treatment. However, stromal reactive phenomena can be modulated with local corticosteroid therapy.

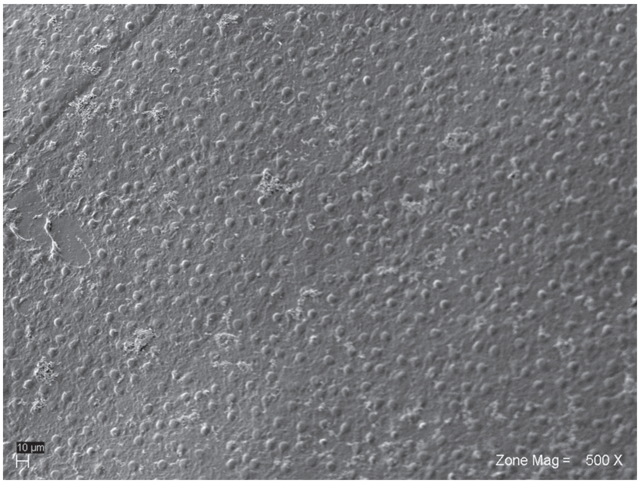


FIGURE 16.3 Electronic microscopy: FSL spots on corneal tissue.

Significant haze after PRK has been described in 3%–11% of all cases [13,14] and in 10–40% of cases with higher optical correction [14–16].

Among the factors influencing the development of a higher haze are high correction levels [13,16,17], small ablation diameters [18], male patients [18], and inappropriate post-operative corticosteroid therapy [13,14].

LASIK complications may be divided in intra- and post-operative complications.

Intraoperative complications are related to malfunctions of the microkeratome (loss of suction, button-hole formation, or free cap) and are partially reduced by FSL application in flap creation.

Dry eye is the most diffuse post-operative complications of LASIK because of corneal nerves interruption occurred during flap creation; normally it is a transient condition, solvable with adequate topical therapy but in case of serious denervation also occlusion of lacrimal punctum by plug is needed [19].

Diffuse lamellar keratitis (DLK) or Sands of Sahara is a non-infectious inflammation under the corneal flap that usually appears between 1 and 3 days following LASIK.

The cause of this inflammation is not known, but it might be a result of immune reactions to antigens within the interface or because of trauma at the time of surgery.

If adequately and early treated, DLK can be treated without causing any significant loss of vision [20–22].

16.4 FSL-Based Procedures

The FSL operates in the infrared range (wavelength: 1053 nm) and uses ultrafast pulses with a duration of 100 fs (10–15 s). Like neodymium-doped yttrium aluminium garnet laser, FSL is solid and capable of causing disruption in stromal tissue through the principle of photoionization, resulting in the rapid formation of a cloud of free electrons and ionized molecules (Figure 16.3). Small volumes of tissue are vaporized, with the formation of cavitation bubbles made up of carbon dioxide and water. This gas is dissipated in the tissue, forming a cleavage plane. First, FSL was approved by the US Food and Drug Administration in 2001 for the creation of corneal lamellae in LASIK.

The fundamental advantage of this innovative technology is related to the fact that the desired tissue was affected while adjacent areas remained intact through an increase in laser pulse frequency and reduction in the amount of energy.

Main application of FSL in refractive surgery is flap creation in LASIK procedure, followed by creation of tunnel for intra-stromal ring implantation, corneal incision for lamellar and penetrating keratoplasty, and astigmatic keratotomy.

Phacoemulsification surgery for incisions, fragmentation of the nucleus, and capsulorhexis represented recent applications of FSL together with corneal refractive lenticule extraction technology (ReLEx) [23].

16.5 ReLEx Procedures

FLEx procedure was introduced in 2007 supported by VisuMax femtosecond laser (Carl Zeiss Meditec, Jena, Germany).

During this surgical technique, the removal of an intra-stromal lenticule, with morphology and thickness consistent with patient myopia, was obtained from under a flap with refractive results were comparable to those observed in LASIK [24].

The necessity to exclude flap creation, and all related complications, led to the development of a new surgical procedure called SMILE.

Small Incision Lenticule Extraction involves the creation of a small 2–4 mm incision, through which a dissector can be introduced to separate the lenticular interfaces and allows the lenticule to be removed through the same incision [25] (Figure 16.4).

The FSL first creates the lower interface of the intrastromal lenticule in a spiral-in pattern (out-to-in direction), followed by a 360° side-cut, followed by creation of the upper interface in a spiral-out pattern (in-to-out direction), known as the cap, and finally, a 2–4 mm access incision (usually superior or superotemporal) that connects the cap interface to the corneal surface [Fig. Laser punti]. Total suction time is approximately 25–35 s (depending on the mode used) and is independent of refractive error treated [25].

By now, scientific literature agreed on satisfying results of SMILE in myopia and myopic astigmatism treatment, also for corneas that already suffered surgery (post-keratoplasty) [26–27].

Adequate stability in corneal biomechanics, absence of flap-related complications (dry eye) and excimer laser complications together with a more rapid and painless visual recovery allow to consider SMILE as a new era of refractive surgery.

16.6 Newer Application of Small-Incision Lenticule Extraction

The availability of a corneal lenticule as a by-product of SMILE, opens up the possibility of using the lenticule for other purposes; in particular, it is under evaluation the possibility to add a stromal lenticule obtained after SMILE for treatment of mild-moderate keratoconus.

In a recent paper ten patients affected by moderate keratoconus were treated with a novel additive surgery called FSL-Assisted Stromal Lenticule Addition Keratoplasty (SLAK) (Figure 16.5). Negative meniscus-shaped stromal lenticules (doughnut-shaped) were produced from corneoscleral eye bank buttons with a refractive lenticule extraction procedure with a 500-kHz VisuMax femtosecond laser.

Recipient corneas underwent a modified FSL flap-cut procedure to produce an intrastromal pocket and lenticules were implanted. Patients were followed up for 6 months after surgery with determination of uncorrected (UDVA) and corrected (CDVA) distance visual acuity, subjective refraction and topographic corneal curvature changes, anterior segment optical coherence tomography (AS-OCT), and in vivo confocal microscopy with acceptable results.

The stromal lenticule addition keratoplasty procedure was clinically efficient in improving the corneal shape and vision in patients with keratoconus. Negative meniscus-shaped lenticule addition induced a flattening of the cone while increasing corneal thickness improving refractive status of these patients [28].

In conclusions, preliminary results are encouraging; however, long-term data on the efficacy of SLAK on keratoconus.

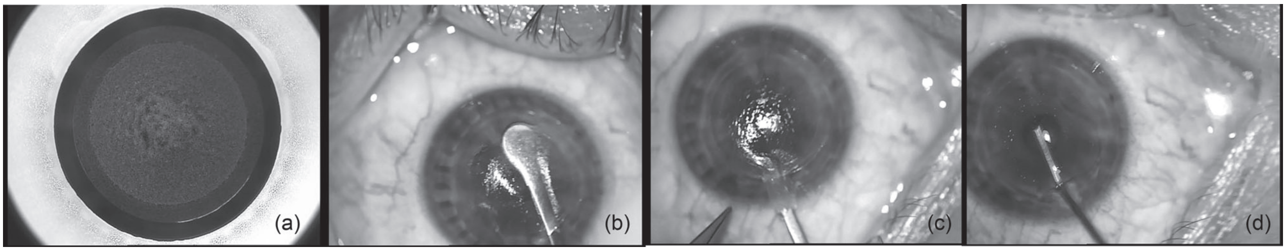


FIGURE 16.4 (a–d) SMILE procedure: lenticue sculpture and extraction.

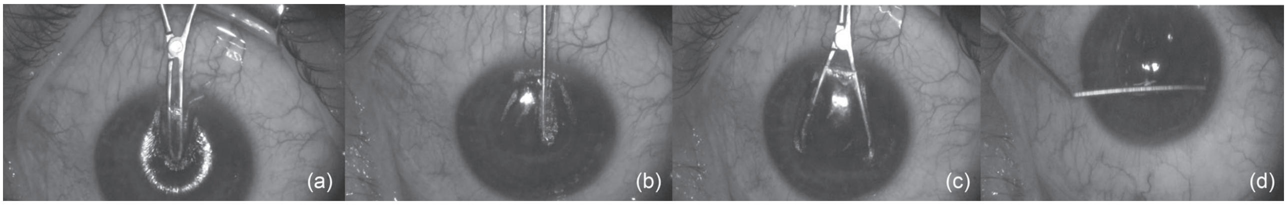


FIGURE 16.5 (a–d) SLAK procedure: lenticule insertion in FSL intrastromal pocket.

REFERENCES

1. Cennamo G, Forte R. 2009. Therapeutic applications: refractive surgery. *Handbook of Laser Technology and Applications*. Taylor and Francis Books, Inc.
2. Sakimoto T, Rosenblatt MI, Azar DT. 2006. Laser eye surgery for refractive errors. Lancet;29:1432–47.
3. McAlinden C. 2012. Corneal refractive surgery: past to present. *Clin Exp Optom*;95:386–98.
4. Wen D, McAlinden C, Flitcroft I, et al. 2017. Postoperative efficacy, predictability, safety, and visual quality of laser corneal refractive surgery: a network meta-analysis. *Am J Ophthalmol*;178:65–78.
5. Kezirian GM, Stonecipher KG. 2004.Comparison of the IntraLase femtosecond laser and mechanical keratomes for laser in situ keratomileusis. *J Cataract Refract Surg*;30:804–11.
6. Reynolds A, Moore JE, Naroo SA et al. 2010. Excimer laser surface ablation – a review. *Clin Exp Ophthalmol*;38:168–82.
7. Donnenfeld ED, O'Brien TP, Solomon R. 2003. Infectious keratitis after Photorefractive Keratectomy. *Ophthalmology*;110:743–7.
8. Pepose JS, Laycock KA, Miller JK et al. 1992. Reactivation of herpes simplex virus by excimer laser photokeratectomy. *Am J Ophthalmol*;114: 45–50.
9. Randleman JB, Caster AI, Banning CS, Stulting RD. 2006. Corneal ectasia after photorefractive keratectomy. *J Cataract Refract Surg*;32:1395–8.
10. Caubert E. 1993. Cause of subepithelial corneal haze over 18 months after photorefractive keratectomy for myopia Refract. *Corneal Surg. (Suppl.)*;9:S65–70.
11. Fantes F, Hanna D, Waring GO III et al. 1990. Wound healing after excimer laser keratomileus (photorefractive keratectomy)in monkeys. *Arch Ophthalmol*;108:665–75.
12. Gartry DS, Kerr Muir M, Marshall J. 1992. Excimer laser photorefractive keratectomy: 18 month follow-up. *Ophthalmology*;99:1209–19.
13. Seilff T, Wollensak J 1991. Myopic photorefractive keratectomy with the excimer laser: one year follow-up. *Ophthalmology*;98:1156–63.
14. Caubert E 1993. Cause of subepithelial corneal haze over 18 months after photorefractive keratectomy for myopia Refract. Corneal Surg. (Suppl.) 9 S65–70
15 Seiler T, Kahle G, Kriegerowski M. 1990. Excimer laser (193 nm) myopic keratomileusis in sighted and blind human eyes. *Refract Corneal Surg*;6:165–73.
16. Heitzman J, Binder P, Kassar B et al. 1993. The correction of high myopia with the excimer laser Arch. *Ophthalmology*;111:1627–34.
17. Seiler T, Wollensak J. 1992. Komplikationen der Laserkeratomileusis mit dem excimerlaser (193 nm) Klin. *Monatsbl. Augenkeilkd*;200:648–53.
18. Yeoh J, Moshegov CN. 2001. Delayed diffuse lamellar keratitis after laser in situ keratomileusis *Clin Exp Ophthalmol*;29:435–7.
19. Calvillo MP, McLaren JW, Hodge DO, et al. 2004. Corneal reinnervation after LASIK: prospective 3-year longitudinal study. *Invest Ophthalmol Vis Sci*;45:3991–96.
20. Hoffman RS, Fine IH, Packer M. 2003. Incidence and outcomes of LASIK with diffuse lamellar keratitis treated with topical and oral corticosteroids. *J Cataract Refract Surg*;29:451–56.
21. DAIZONG WEN, COLM MCALINDEN, IAN FLITCROFT et al. 2017. Postoperative Efficacy, Predictability, Safety, and Visual Quality of Laser Corneal Refractive Surgery: A Network Meta-analysis Am J Ophthalmol;178: 65–78.
22. Chung SH, Mazur E. 2009. Surgical applications of femtosecond laser. *J Biophotonics*;2:557–72.
23. Callou TP, Garcia R, Mukai A et al. 2016. Advances in femtosecond laser technology. *Clin Ophthalmol*;10:697–703.
24. Vestergaard A, Ivarsen A, Asp S, Hjortdal JØ. 2013. Femtosecond (FS) laser vision correction procedure for moderate to high myopia: a prospective study of reLEx® flex and comparison with a retrospective study of FSlaser in situ keratomileusis. *Acta Ophthalmol*;91:355–62.
25. Ganesh S, Brar S, Arra RR. 2018. Refractive lenticule extraction small incision lenticule extraction: a new refractive surgery paradigm. *Indian J Ophthalmol*;66:10–9.
26. Alió Del Barrio JL, Vargas V, Al-Shymali O. 2017.Small incision lenticule extraction (SMILE) in the correction of myopic astigmatism: outcomes and limitations - an update. *Eye Vis (Lond)*;15(4):26.
27. Mastropasqua L, Calienno R, Lanzini M, Nubile M. 2015. Small incision lenticule extraction after deep anterior lamellar keratoplasty. *J Refract Surg*;31:634–7.
28. Mastropasqua L, Nubile M, Salgari N, Mastropasqua R. 2018. Femtosecond laser-assisted stromal lenticule addition keratoplasty for the treatment of advanced keratoconus: a preliminary study. *J Refract Surg*;34(1):36–44.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

17

Photodynamic Therapy

Brian C. Wilson and Stephen G. Bown

CONTENTS

17.1 Introduction 249
17.2 Light Sources 250
  17.2.1 Lasers 250
  17.2.2 Non-laser Sources 252
17.3 Light Delivery Systems 252
17.4 Optical Monitoring and Dosimetry 255
17.5 Alternative Photosensitizer Activation Schemes 257
  17.5.1 Two-photon Activation 257
  17.5.2 Up-converting Nanoparticles 257
  17.5.3 Bioluminescence Activation 257
  17.5.4 Ultrasound Activation 258
  17.5.5 Ionizing Radiation Activation 258
17.6 PDT for Cancer Treatment 258
17.7 Other Applications of PDT 259
17.8 Conclusions 259
Bibliography 260

17.1 Introduction

Photodynamic therapy (PDT) is the use of drugs (photosensitizers) that are activated by light, generating photoproducts that kill or modify cells and tissues, usually in the presence of oxygen depending on the photosensitizer and the photophysical pathways involved. While the basic concept is over a century old, its development was significantly accelerated by the emergence of component technologies: lasers and other practical high-brightness light sources such as light emitting diodes (LEDs); optical fibres that enable efficient delivery of the light to many sites within the body and synthesis of photosensitizers with high activation efficiency (photoproduct quantum yield, good bio-distribution and disease targeting) and different biological mechanisms of action. For most photosensitizers currently in clinical use or under investigation, the active photoproduct is believed to be mainly an excited singlet state of molecular oxygen ($^1O_2$) that is generated by the photophysical process illustrated in Figure 17.1. In this process, the photosensitizer returns to its ground state after energy transfer to ground-state molecular oxygen ($^3O_2$) and can be re-activated by further photon absorption Thus, it acts essentially as a catalyst for oxygen activation. The active photoproducts damage tissue/cell components, primarily membranes, for example, through lipid peroxidation, leading to cell death or altered function. Particularly in a low-oxygen environment or with some photosensitizers even in the presence of free oxygen, other processes such as direct interactions of the excited singlet or triplet-state molecule with biomolecules may also contribute or dominate. Absorption (activation) spectra for representative PDT photosensitizers are shown in Figure 17.2. Note that these typically show multiple absorption peaks across a range of wavelengths. Most often, the longest wavelength peak is used for clinical applications, since this gives the greatest penetration of the light into tissues. However, increasing the

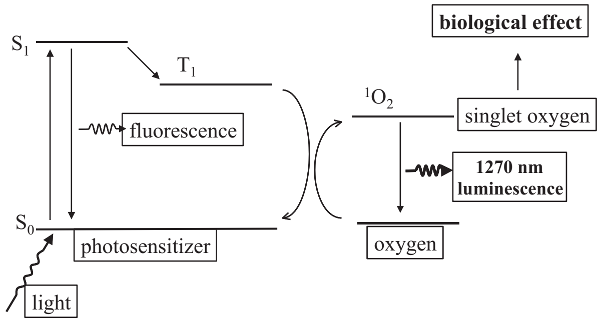


FIGURE 17.1 Energy diagram for singlet oxygen ($^1O_2$) generation in (Type II) activation of photosensitizers. An excited singlet state of the photosensitizer, S_1, is produced following absorption of a photon by the ground state, S_0. This undergoes inter-system crossing to a long-lived triplet state, T_1, which then transfers energy to triplet ground-state molecular oxygen ($^3O_2$), generating singlet oxygen, $^1O_2$. S_1 may also de-excite radiatively producing fluorescence or non-radiatively, while $^1O_2$ may decay back to the ground state, emitting a NIR photon.

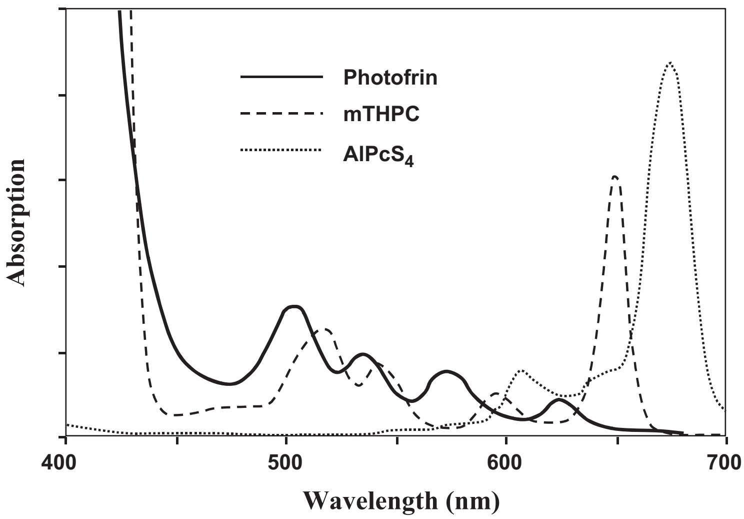


FIGURE 17.2 Typical absorption (activation) spectra for 3 different photosensitizers. Note the large absorption peaks in the far-red region with second-generation mTHPC and AlPcS$_4$ compared to the small red peak for the first-generation photosensitizer Photofrin® (sodium porfimer).

activation wavelength beyond about 800 nm is not useful, since water absorption starts to limit the light penetration and the quantum energy becomes insufficient to efficiently generate photoproducts.

PDT is a two-step process. First, the photosensitizer is administered, usually by intravenous injection or, for skin diseases, by topical application. A time interval that varies between minutes to days depending on the photosensitizer pharmacokinetics is then allowed for accumulation in the target tissue or organism. The light is then applied to the local site of disease. Currently, several different PDT drugs are approved for use for treatment of various solid tumours and pre-malignant tissues, sun-damaged skin, age-related macular degeneration (a retinal vascular disease that leads to blindness) and local bacterial infections. Clinical and other pre-clinical trials have been carried out or are in progress with several other photosensitizers, for applications such as treatment of fungal infections, coronary vessel disease and rheumatoid arthritis, among others. PDT for treating tumours is discussed in more detail below to illustrate the versatility and some of the practical issues in the technique.

There are many considerations in optimizing the delivery of PDT. These relate to the properties of the photosensitizer, the cell and tissue photobiological responses (that include direct cell kill, vascular damage and immune effects), and the optical technologies. Here, we will address the last of these in detail: for a wider overview, the reader is referred to the bibliography at the end. Three groups of optical technologies will be considered: light sources, fibre-optic delivery systems and optical methods for treatment monitoring. In each case, the current status, technological limitations and potential future needs will be discussed.

17.2 Light Sources

Light sources for clinical PDT must provide enough light of an appropriate wavelength (range) for efficient photosensitizer activation, in an acceptable treatment time and with a spatial distribution that matches the target tissue geometry. Safety, reliability (constancy of output power and wavelength), ease-of-use, low maintenance, use of standard electrical power, size and (trans)portability, lifetime and cost are other significant practical considerations. Governmental approvals are required for clinical devices such as PDT light sources, in addition to the photosensitizers or in some cases the specific combination. This necessitates incorporating optical, electrical and mechanical safety features as well as built-in and/or external output-monitoring tools that are not generally found in laboratory or commercial laser systems. With some exceptions in which only very superficial disease is involved, wavelengths in the range ~630–800 nm are required in order to have adequate depth penetration in tissue, as illustrated in Figure 17.3, and adequate quantum energy to generate a triplet state for efficient $^1O_2$ production. Typically, delivered optical power of up to a few Watts is required with current photosensitizers, depending on the area or volume of tissue involved. Three main types of light source have been used for both pre-clinical research and clinical treatments: continuous wave (CW) diode lasers, LED arrays and, less commonly, filtered high-brightness lamps.

17.2.1 Lasers

The main reason to use lasers for PDT is to achieve efficient coupling of light into optical fibres for light delivery into the body, either endoscopically or interstitially, i.e. with fibres inserted directly into tissue through needles or catheters positioned under some form of image guidance. Historically, argon ion-pumped dye lasers were the standard clinical light source for PDT until about 1990, with pump powers of 5–20 W and red output power of typically 1–4 W. These had the advantage of wavelength tunability, allowing a single system to be used for a range of different photosensitizers with different activation wavelengths. However, these lasers were expensive and difficult to use and maintain in the clinical environment, even when re-engineered from laboratory into clinical systems.

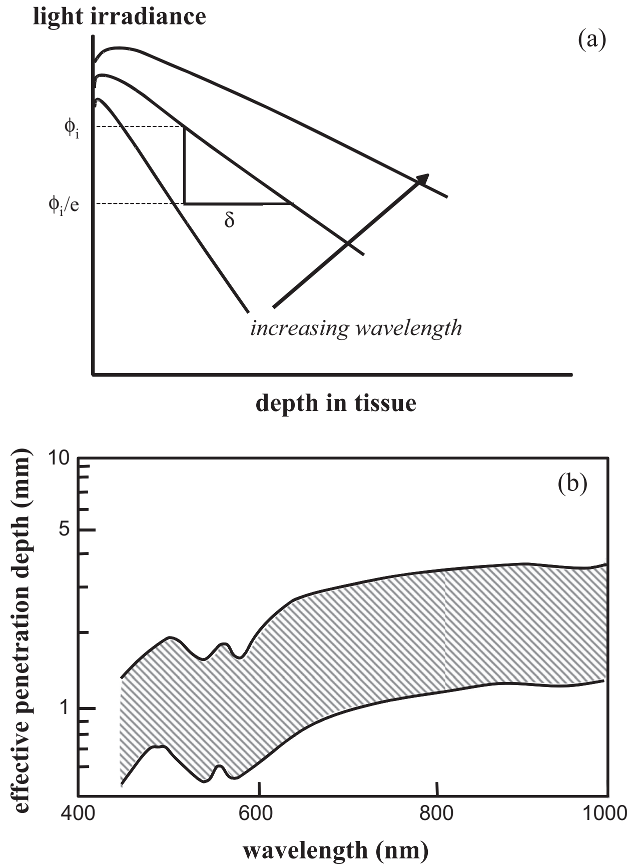


FIGURE 17.3 Penetration of light in tissue. (a) Schematic of fluence-depth curves in tissue irradiated by a broad light beam at wavelengths where the tissue is both scattering and absorbing. This also shows the operational definition of the effective penetration depth, *δ*. (b) Wavelength dependence of *δ* in the visible-NIR spectral range. The upper and lower curves represent slightly- and highly pigmented tissues, such as the brain and liver, respectively. Other tissues are intermediate but show the same general shape. The rise above 600 nm is due mainly to the decrease in blood (haemoglobin) absorption. For further details on the tissue optical properties and experimental and theoretical studies of light propagation in the tissue, see Chapter by Jacques and Patterson in this volume.

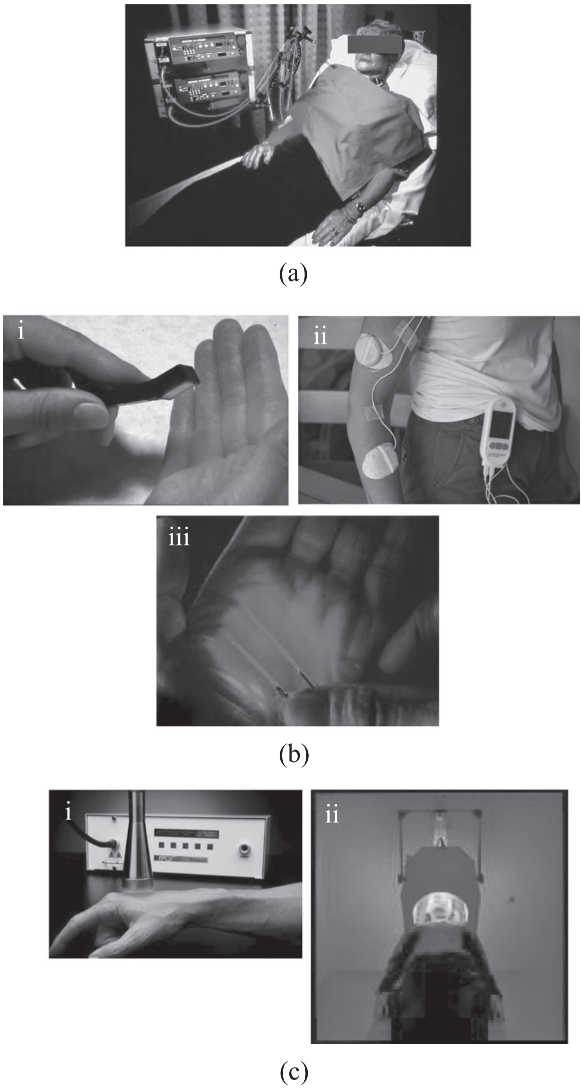


FIGURE 17.4 Light sources used in PDT. (a) Diode laser system providing several Watts of red-/NIR light into multi-mode optical fibres. (b) LED arrays: (i) small planar array for intra-oral irradiation (courtesy: PRP Optoelectronics, UK). (ii) Flexible wearable LED array sources for skin treatment, showing also the power supply carried on the belt (courtesy: Ambicare, UK). (iii) Interstitial linear arrays (courtesy: Light Sciences, USA). (c) Filtered lamps: (i) arc lamp with fibre-bundle delivery (courtesy EXFO, Canada). (ii) Array of fluorescent tubes in U shape for head and face treatments—note the blue filtering to limit the penetration for superficial pre-malignant skin treatments (courtesy DUSA, USA).

The argon pump laser was superseded by frequency-doubled Nd:YAG (using a potassium titanyl phosphate, KTP, crystal) lasers in the late 1980s, markedly improving the system reliability. Although pulsed, the repetition rate of KTP lasers is high enough (~10 kHz) that there is no significant saturation of the photosensitizer excited state that can occur with shorter-pulse, high peak power activation. Hence, the KTP laser was shown to be equivalent in terms of photosensitizer activation and photobiological responses to a CW laser of the same average power. However, high cost and bulkiness remained as limitations. Since the mid-1990s, diode lasers have become the standard laser device in PDT: a typical system is shown in Figure 17.4a. Typically, these have a CW output power of 2–4 W, with 80%–90% coupling efficiency into 200–400 μm diameter multi-mode optical fibres. The major limitations of current clinical laser diodes systems are that (a) the cost per Watt remains quite high (~$10K), especially since (b) each system provides only a single wavelength, which must be selected for a specific photosensitizer so that multiple units may be required, and (c) they are not readily available for all wavelengths of interest, especially in the 700–800 nm range.

It should be noted that, although these lasers use essentially standard electro-optic components, the cost is much higher for a clinical system than for a laboratory-based source of the same output. This is due to the need to package the technologies into a unit suitable for the clinical environment, additional safety features such as interlocks, the cost of governmental approvals and the incorporation of additional built-in output monitoring and calibration devices. Several commercial diode lasers approved for clinical PDT are available.

An emerging variant of PDT that could necessitate other types of laser is 2-photon activation. This can be done in two

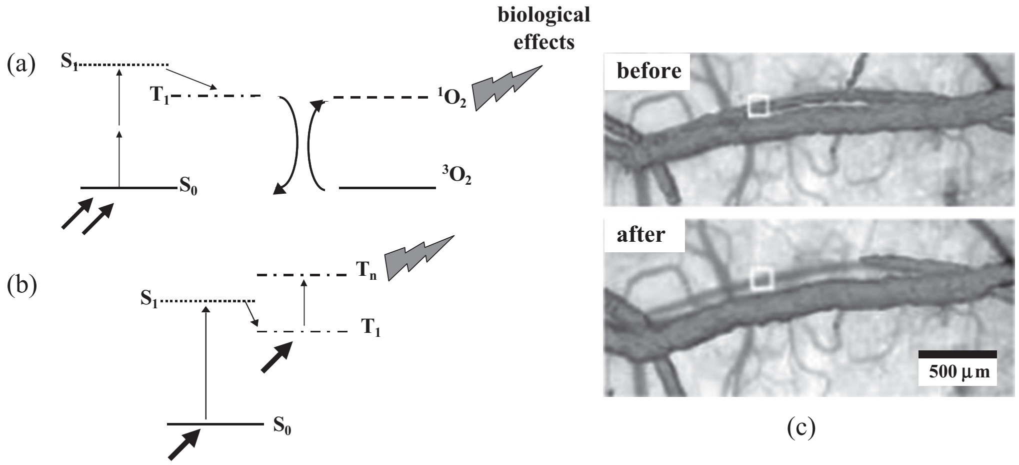


FIGURE 17.5 2-photon PDT. (a) Energy diagram for simultaneous 2-γ activation, leading to the generation of singlet oxygen. (b) 2-γ, 2-λ activation leading to biologically-active excited triplet states of the photosensitizer. (c) Example of closure of a single blood vessel by simultaneous 2-γ PDT, showing the high precision of the spatial localization of the PDT effect: the treatment spot is indicated by the white square.

different ways, as illustrated in Figure 17.5a and b: simultaneous 2-photon [2-γ] and 2-photon/2-colour [2-γ, 2-λ] activation. The rationale and current status of these are discussed in more detail below. From the light-source perspective, they require different laser technologies. In [2-γ]-PDT a pulsed laser is used, such as a Ti:sapphire amplifier seeded by a femtosecond oscillator, producing typically 300 fs pulses at around 80 MHz pulse repetition frequency and average powers of tens of mW. In [2-γ, 2-λ]-PDT activation by an initial pulse at a wavelength corresponding to the ground-state absorption spectrum of the photosensitizer leads to the formation of the triplet state. A second pulse (e.g. at a wavelength that is absorbed by the triplet state) generates higher energy excited states that then go on to produce photochemical damage. This does not require extremely high peak-power densities nor ultrashort pulses as for simultaneous [2-γ] activation: e.g. initial studies used a frequency-doubled Nd:YAG laser (~10 ns) as the first source and a flash lamp-pumped dye laser (~400 ns) as the second source. The laser pulses must be synchronized to match the inter-state transition kinetics of the photosensitizer.

17.2.2 Non-laser Sources

Before discussing the status of non-laser light sources for PDT, namely, LEDs and lamps, an important consideration in the use of broad-band sources is the overlap between the source output spectrum and the photosensitizer activation spectrum, as illustrated in Figure 17.6. Even if these fully coincide (i.e. have the same peak wavelength and bandwidth), the effective source power for a typical 25 nm bandwidth LED is only about half of that of a laser of the same optical power at the peak photosensitizer wavelength. Hence, one typically needs about twice the output power with LEDs or lamps as with a laser to achieve the same $^1O_2$ yield.

LEDs have been configured as planar arrays of different sizes and format, as illustrated in Figure 17.4b. Typically, these deliver power densities of up to about 100 mWcm^{-2}. Another approach is to use a linear array of LEDs within a small diameter catheter for interstitial use. To date these have low output (a few mW per cm length) to avoid generating too much heat in the tissue, but can be implanted to allow irradiation over a period of hours rather than minutes, which may induce different photobiological mechanisms of action. A possible future development is to incorporate active cooling to allow higher powers. LEDs have also been configured into "wearable" devices, where long treatment time compensates for the low output power. The advantage lies in the flexibility that allows the source to be shaped to match curved tissue surfaces. Organic LEDs have also been investigated for this purpose.

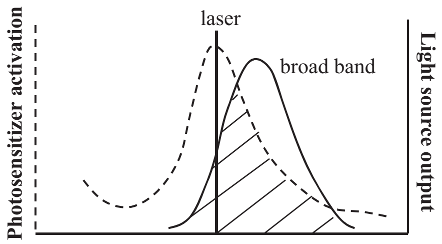


FIGURE 17.6 Spectral overlap between a broad-band light source and photosensitizer action spectrum. The effective irradiance is the convolution of the 2 spectra.

High brightness arc or fluorescent lamps, with external filtering to remove the infrared (avoiding tissue heating) and to select the most appropriate excitation band, are also in use. The former can be coupled adequately into optical fibre bundles or liquid light guides, or even into large diameter (~mm) single fibres for delivery. Fluorescent lamps have been configured to irradiate large surface areas externally.

17.3 Light Delivery Systems

The purpose of light delivery systems for PDT is to transport the light from the source to the target tissue with high efficiency and to distribute it such that it matches the 2- or 3-D geometry of the tissue. The simplest situation is the treatment of accessible tissues where the penetration depth of the light is adequate to irradiate the diseased tissue to the required dose, such as in treating the skin or exposed tissues during surgery. In some cases, no additional delivery system is required and the treatment field is defined by the emission distribution of the source itself, as in the planar LED arrays and the fluorescent tube array shown in Figure 17.4b and c. With laser sources, it is possible to use a beam expander directly to obtain the required treatment area. However, a more general approach is to deliver the light using large-diameter multi-mode optical fibres coupled to the laser, as mentioned above, giving >90% coupling efficiency. It is necessary to make the output flat (say, ± 15%) across the treatment field, and this can be done with a simple lens system, including a microlens fixed to the fibre itself.

For endoscopic applications to treat inside hollow organs such as the lung, gastrointestinal tract and bladder, fibre-optic light delivery is required and the fibre is usually placed within the instrument channel of the endoscope, as illustrated in Figure 17.7a or via a special applicator as in Figure 17.7b. For treating large tissue volumes (thicker than the penetration depth of the light), the light must be delivered interstitially, i.e. into the tissue. A simple cut-end fibre used for this purpose will produce a roughly spherical irradiation zone due to high light scattering of tissue at visible/near-infrared (NIR) wavelengths. Diffusing-tip fibres such as those shown in Figure 17.7a and b have been developed that produce a cylinder of light up to a few cm in length with high uniformity (Figure 17.7c). The standard means to do this is to strip the fibre-cladding along the required length and then coat the surface with a highly light-scattering material. A recent alternative is to write a Bragg grating into a (single-mode) fibre, thereby modifying the refractive index to produce light-scattering centres along the length rather than to provide wavelength filtering as used in optical telecommunication. The specific advantage of this approach is that it provides very fine optical fibres that can have very long diffusing regions, suitable for either intra-vascular applications or to irradiate long segments of other hollow structures without needing to move the tip.

In interstitial treatment of large volumes, the spacing of diffusing fibres is critical in order to ensure complete coverage, i.e. to achieve adequate local fluence (Jcm^{-2}) at all points in the tissue. This is strongly dependent on the optical absorption and scattering properties of tissue at the treatment wavelength. Figure 17.7d shows the distribution for a single ideal linear diffusing fibre in a specific absorbing/scattering medium, while Figure 17.8a shows multiple interstitial fibre deliveries into the prostate and the corresponding fluence distributions, based on finite-element numerical solutions of the photon transport. For treatments that require multiple optical fibre sources, multiple laser sources must be used or the light from a single laser of higher power must be split efficiently between the fibres. Both open-beam and fibre-fibre coupling methods have been demonstrated for this purpose. It should be noted that these devices need to handle much higher powers than is typical in fibre-optic telecommunications. Variable-power instruments have also been demonstrated, as have instruments that incorporate feed-back control based on local light fluence-rate measurements in tissue: an example is shown in Figure 17.8b.

Other ways of distributing the light throughout the tumour volume or over a large surface area include balloon applicators (Figure 17.9a and b) and fibre-optic "textile" (Figure 17.9c). The former are used to give uniform illumination within a body cavity, such as in the oesophagus, cervix, bladder pleural cavity or abdominal cavity, or following tumour resection (e.g. in the brain). In this case, fibre-optic delivery from a diode laser source takes the light into (or close to) the centre of the inflatable balloon, which may be simply spherical or shaped to conform to the cavity, for example, a cylindrical shape with a central linear diffusing fibre. The balloon is inflated to have good contact with the tissue surface (without applying pressure that could reduce blood flow and hence oxygenation that

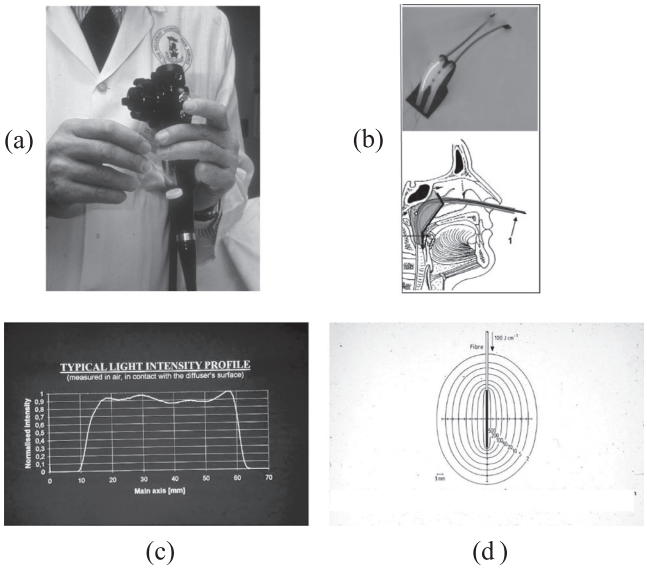


FIGURE 17.7 Linear diffusing fibres for light delivery in PDT. (a) Diffusing-tip fibre being placed within the instrument channel of an endoscope to treat lesions in the oesophagus. (b) Two diffusing fibres integrated into an applicator for treating nasopharyngeal tumours. (c) Output of a diffusing optical fibre showing uniformity along its length. (d) Iso-fluence contours from linear diffusing fibre in tissue-like absorbing and scattering medium.

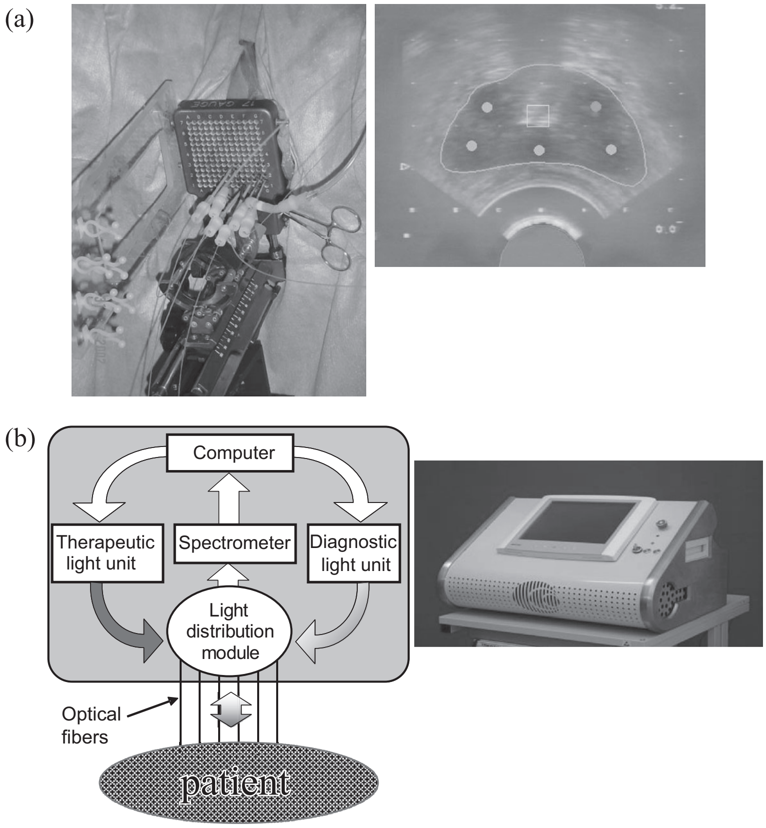

(a)
(b)
Computer
Therapeutic light unit
Spectrometer
Diagnostic light unit
Light distribution module
Optical fibers
patient

FIGURE 17.8 Multiple interstitial fibre treatment of prostate cancer. (a) Set-up showing the treatment fibres and light detector probes placed through a template that is co-registered with an ultrasound probe placed in the rectum to image the tissue and the fibres. Also shown is the corresponding light fluence distribution for a 5-source arrangement in the prostate, calculated by a finite-element solution of diffusion theory for light propagation in a scattering and absorbing medium. The prostate, urethra (square) and rectum (semi-circle) are marked on an ultrasound cross-sectional image. (b) Schematic (after Swartling et al., 2010) of a computer-controlled light delivery and dosimetry system for multiple-fibre treatments (courtesy: SpectraCure, Sweden).

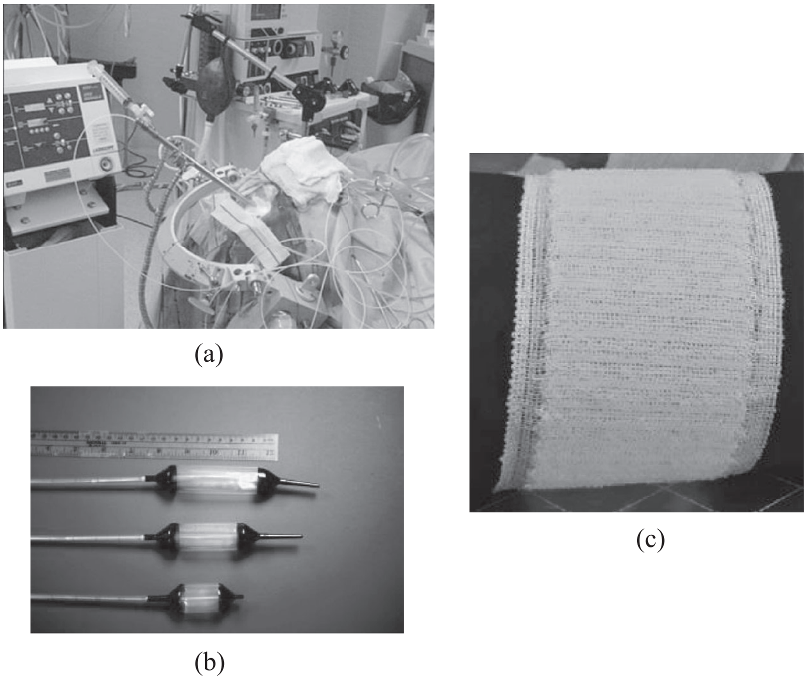


FIGURE 17.9 (a, b) Balloon applicators used to distribute PDT treatment light within body cavities: (a) spherical balloon used in the brain after surgical removal of bulk tumour; (b) applicators for use in the oesophagus; (c) light "textile" for large area treatments (after Cochrane et al., 2013).

is needed for PDT activation). To obtain uniform light distribution across the whole surface, the balloon may be made of a light scattering material, or a diffusing fibre can be used together with saline or air for inflation, or the balloon can be filled with an optically turbid fluid, most commonly intralipid, a fat-protein colloid. Fibre-optic cloth, or textile, was introduced recently in which plastic optical fibres are woven or knitted to form a flexible sheet that can be placed into complex body cavities. Laser light is injected into both ends of the individual fibres, and the leakage from the fibres distributes the light uniformly (±~15%) across the area. This approach is particularly useful for large surface areas such as in the chest, abdomen or pelvis.

Monitoring of the local fluence rate at critical points in the tissue can be done also using implanted optical fibre probes connected to a sensitive photodetectors: some of the fibres seen in Figure 17.8a are used in this way. Short-length, diffusing-tip fibres have a near-isotropic response that, with proper calibration, gives the true local fluence rate independent of the local radiance distribution, i.e. of the angular dependence of the light field. An alternative is to use the same cut-end fibres for light delivery and detection, switching between these to distribute and monitor the light throughout the tumour volume: this is done in the system shown in Figure 17.8b

The final effect of the PDT treatment on the tissue depends not only on the local light fluence but also on the photosensitizer distribution and the tissue oxygenation. These can be measured within the tissue using fibre probes based on fluorescence and diffuse reflectance spectroscopy, respectively: in the system shown in Figure 17.8b, these functions are built into a single instrument with the detected light from the various interstitial fibres being switched to corresponding spectrometers.

For the future, one potential approach to light delivery specifically in PDT of retinal disease is to use a scanning laser beam. This could be implemented, for example, by coupling the therapeutic beam into an imaging laser-scanning confocal ophthalmoscope. In this case, a high-resolution image would first be made using one or more low-power lasers at the appropriate wavelengths to image the disease (e.g. abnormal vasculature). The region to be treated would then be defined on this image, following which the system would be switched over to the focused treatment laser that would be scanned to irradiate this region or to trace out individual blood vessels. At this time, no studies have been done on the PDT response of tissues using a focused scanning beam, i.e. performing point-by-point irradiation, so that the optimum irradiation parameters have not been defined. For example, will the total fluence required be similar to that for whole-field illumination? Given the relatively poor mode structure of diode lasers, it may also be difficult to couple these sources to provide a highly focused beam of adequate power into the ophthalmoscope, so that other laser technologies may be needed.

17.4 Optical Monitoring and Dosimetry

As indicated above, online measurement of PDT treatments is a critical aspect that involves a variety of optical techniques and technologies. These include:

- the use of optical fibre probes with modified tips to provide near-isotropic response for measuring light irradiances on or in tissues,
- fluorescence or diffuse reflectance spectroscopy to measure the concentration of photosensitizer non-invasively *in vivo*,
- optical (and non-optical) probes for measuring tissue oxygenation before, during and after PDT treatment.

In addition, several optical diagnostic methods are also of potential value for monitoring the biophysical and/or biological changes that result from treatment. For example: laser Doppler imaging has been used to image the overall changes in tissue blood flow after PDT; Optical Coherence Tomography, the optical analogue of high-frequency ultrasound that is based on light interferometry using short (~100 nm) coherence length sources such as superluminescent diodes, can detect changes in the tissue structure or the presence of edema. The Doppler version of this, based on detecting single-scatter ("echoes") from moving red blood cells, has been shown capable of similarly tracking the changes in blood flow within microvasculature *in vivo*, including during PDT irradiation. Photoacoustic imaging, in which a short laser pulse is used to generate an acoustic signal due to local transient energy absorption within the tissue, has also been used for PDT response monitoring. In particular, based on haemoglobin light absorption, the sharp boundary in the blood concentration in the tissue can be localized following vascular shut-down in the treated region. This can then be used to define the effective treatment in depth and volume.

Rather than providing an exhaustive summary of these various methods, the focus here will be on one specific, leading-edge technique (SOLD—Singlet Oxygen Luminescence Dosimetry) that involves the use of a laser source and recently developed high-sensitivity photodetectors for measuring directly the generation of ${}^1O_2$ in tissue.

Recalling Figure 17.1, after ${}^1O_2$ is generated it may interact with bio-molecules or decay back to the oxygen ground state by the emission of a photon at 1270 nm. This is a quantum-mechanically forbidden transition, so that the signal is intrinsically weak. Nevertheless, this is a standard ${}^1O_2$ measurement technique in photochemistry, e.g. in studies of photosensitizers in solution to determine the ${}^1O_2$ quantum yield. However, since ${}^1O_2$ is highly reactive, its lifetime is greatly shortened in cells or tissues (<~200 ns) compared to that in non-biological media or to its natural radiative lifetime (~4s). As shown in Figure 17.10d, the luminescence signal first increases following a short light pulse (<<triplet-state and single oxygen lifetimes), as the T1 state is populated and transfers energy to ${}^3O_2$, and then decays. If the triplet-state lifetime is much greater than the singlet-state lifetime, which is usually the case, then the singlet oxygen generation rate is given by

$$\left[{}^1\mathrm{O}_2\right](t) = \frac{N\sigma[S_0]\Phi_D\tau_D\left(\exp(-t/\tau_T) - \exp(-t/\tau_D)\right)}{\tau_T - \tau_D} \tag{17.1a}$$

where [ ] represents the transient state concentrations, N is the number of photons per cm^2 incident on the sample during the

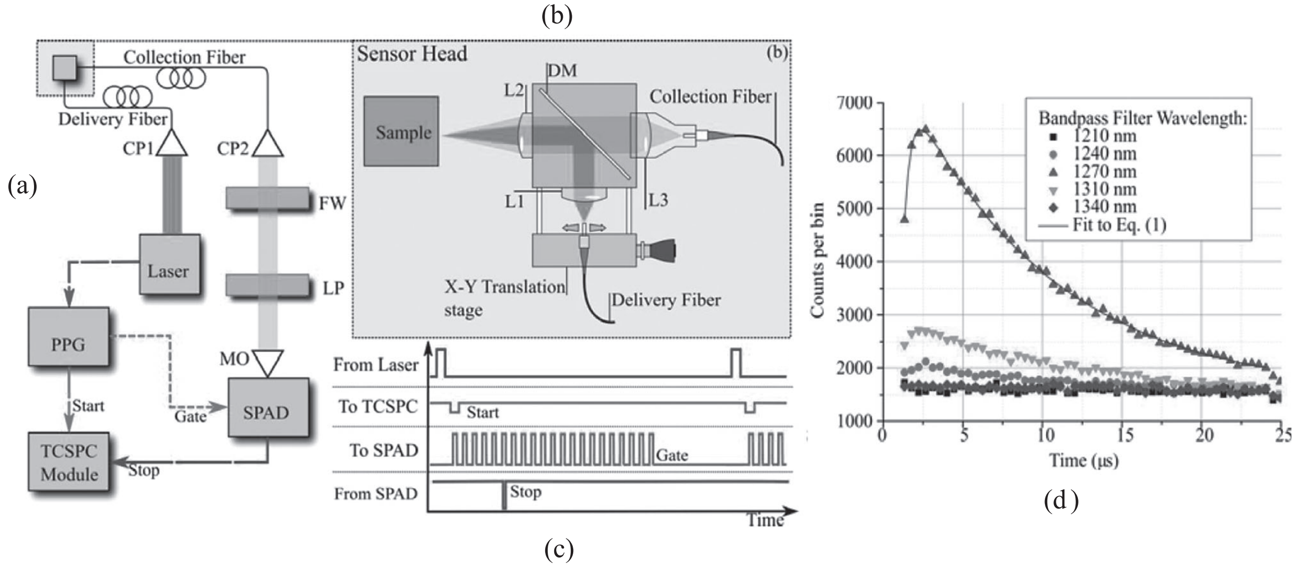


FIGURE 17.10 $^1O_2$ NIR luminescence detection (after Gemmell et al., 2017). (a) Schematic of a fibre-optic probe instrument, based on a SPAD time-correlated single-photon detector. (b) Details of the probe head. (c) Pulse train. (d) Example of time spectra at different wavelengths for photosensitizer solution, showing the high signal at the singlet-oxygen emission peak at 1270 nm. CP, collimator package; DM, dichroic mirror; FW, filter wheel; L, lens; LP, long-pass filter; MO, microscope objective; PPG, pulse pattern generator; TCSPC, time-correlated single-photon counter.

pulse, σ is the photosensitizer ground state absorption cross section [cm^2], Φ_D is the $^1O_2$ quantum yield, τ_T is the photosensitizer triplet-state lifetime and τ_D is the $^1O_2$ lifetime. The singlet oxygen luminescence emission [photons cm^{-3} s^{-1}] at time t is then

$$L_{1270}(t) = \frac{[^1O_2](t)}{\tau_R} \tag{17.1b}$$

where τ_R is the $^1O_2$ radiative lifetime in the specific microenvironment. Hence, the total time-integrated signal is

$$L_{1270}(t)dt = \frac{N\sigma[S_0]\Phi_D\tau_D}{\tau_R} \tag{17.1c}$$

The total number of luminescent photons is, therefore, proportional to the $^1O_2$ lifetime, making the signal correspondingly much weaker in cells and tissues than in solution. In addition, until recently, detectors sensitive at 1270 nm (e.g. germanium crystals) had low responsivity and long response times, so that even time-integrated detection in true biological media was marginal or impossible.

A technological breakthrough in the 1990s was the development of a fast, high-sensitivity photo-multiplier tube (Hamamatsu Corp) with extended NIR response for optical telecommunications at 1.3 μm. More recent developments have been in superconducting nanowire single-photon detectors and semiconductor-based single-photon avalanche diodes (SPADs) that have at least an order of magnitude greater quantum efficiency than photomultiplier tubes. $^1O_2$ luminescence from photosensitizers in solution and tissue-simulating phantoms has been reported with these detectors, and a particular advantage is that their small size and high quantum efficiency enable them to be incorporated into a fibre-optic-based probe rather than in only open-beam configuration, which makes for improved compatibility, flexibility and practicality for *in vivo* dosimetry. This is also illustrated in Figure 17.10.

The laser pulse length required for single-photon counting to directly measure the signal kinetics in $^1O_2$ luminescence dosimetry is around 10 ns to capture the fast rise time of the signal. Initially, an optical parametric oscillator pumped by a frequency-doubled ND:YAG laser was used, providing 10 ns pulses. The limitation was its low repetition rate (10 Hz) and low average power (~10 mW), requiring long signal integration times to obtain adequate signal-to-noise ratio. The source used currently in most reports is a compact frequency-doubled Nd:YLF laser (532 nm, 10 ns pulses, up to 10 kHz pulse repetition rate, up to ~ 500 mW average power). Possible future developments will include using this laser to pump a dye cell in order to have a tunable source that can be matched to the absorption maxima of different photosensitizers, and incorporation of a beam-scanning system to enable $^1O_2$ imaging in tissue. There have also been reports using a chopped CW diode laser as the activation source for SOLD. Clearly, the ‘pulses’ cannot then be in the ns range since external beam chopping or electronic pulsing is not fast enough. Rather, the idea is to utilize the trailing edge of the much broader square-wave and gate the photodetector accordingly. The potential limitation of this approach is that, for the same average optical power delivered to the tissue, the fraction that contributes to generating the measured signal is very small, most of the light within the square wave being “wasted”. The advantage is that a much wider range of wavelengths is available at relatively low cost, and the same beam can be used both for SOLD and for the actual PDT treatment.

17.5 Alternative Photosensitizer Activation Schemes

This section will focus on alternative ways to activate PDT photosensitizer that are different from "conventional" single-photon (linear) activation with diode lasers, LEDs and lamp sources. These alternatives include both laser-based and non-photonic approaches.

17.5.1 Two-photon Activation

Both simultaneous 2-photon and 2-photon/2-color activation of photosensitizers were introduced above in terms of the laser sources needed. In the first approach, the S_1 state of the photosensitizer ($E_1 \propto 1/\lambda_1$) is generated by simultaneous absorption of two photons each of wavelength $2\lambda_1$. This confers two putative advantages. First, the source wavelength ($2\lambda_1$) is typically in the NIR range, so that the light penetration in tissue is higher than with single-photon activation in the visible range. At the same time, the activation wavelength, λ_1, gives efficient photoproduct generation. The limitations are as follows. The rate of energy absorption and hence rate of photoproduct generation, R, for 2-photon PDT depends quadratically on the local fluence rate, φ ($2\lambda_1$, d), at any given depth, d, in tissue:

$$R_{2-\text{photon}}(d) \propto \gamma(\lambda_1) C\,\sigma\,\varphi(2\lambda_1,d)^2 \tag{17.2a}$$

$$\varphi(2\lambda_1,d) \propto \gamma(\lambda_1)\,C\,\sigma\,\exp - 2\mu_{\text{eff}}(2\lambda_1)/d \tag{17.2b}$$

where μ_{eff} is the effective attenuation coefficient in tissue (which depends on the light absorption and elastic scattering). This should be compared with the situation for single-photon activation:

$$\varphi(\lambda_1,d) \propto \gamma(\lambda_1)\,C\,\varepsilon(\lambda_1)\exp\{-\mu_{eff}(\lambda_1)/d \tag{17.3}$$

where ε is the absorption coefficient per unit photosensitizer concentration. Hence, 2-photon activation will only give larger treatment depth in tissue if (i) the photosensitizer has high enough 2-photon cross-section and (ii) the optical attenuation coefficient of the tissue at wavelength $2\lambda_1$ is less than half the attenuation coefficient at $2\lambda_1$ (to compensate for the additional factor of 2 in the exponent). Standard PDT photosensitizers typically have very low 2-photon cross-section ($\sigma <\sim 10$ Goppert-Meyer units), but compounds specifically designed to have high 2-photon cross-section ($\sigma > 1000$ Goppert-Meyer units) at long activation wavelengths have been reported and have demonstrated biological responses in both cell culture *in vitro* and *in vivo* animal models. The situation with respect to point (ii) is not so clear: as seen in Figure 17.3b the light penetration (which is inversely proportional to the effective attenuation coefficient) tends to level off above about 650–700 nm and so usually does not achieve the 2-fold increase required. Tumour responses in mice with up to 2 cm overlying tissue have been reported, but interpretation of these results is controversial, since other factors may be involved.

A second advantage of simultaneous 2-photon PDT also arises from the same phenomenon as described by Equation (17.1a). Using a focused near-IR fs laser beam, it is possible to activate the photosensitizer only within a small focal volume where the light intensity is high enough to overcompensate for the limited 2-photon cross section. This is exploited in 2-photon confocal fluorescence microscopy to achieve larger depth of imaging (up to ~ 100 µm, which is about double that of standard 1-photon confocal microscopy) without causing photobleaching of the fluorophore or photodamage to the specimen. Note that the necessary high power density of the laser beam at the focal spot is achieved without corresponding high local energy density, through the use of the ultrashort laser pulses. This results in "ultra-precise" PDT, as illustrated in Figure 17.5c, where simultaneous 2-photon activation was used to shut down single blood vessels without affecting overlying or adjacent tissues. Since the laser beam is tightly focused to achieve good 2-photon efficiency, the maximum power that can be used is limited by the need to stay below the threshold for photomechanical damage (~10^{12} W cm^{-2}). One potential application is to target the so-called "feeder vessels" to destroy the downstream abnormal vasculature in macular degeneration of the retina, which is the leading cause of blindness in the elderly. This is still at the pre-clinical research stage.

Putting aside the challenge of the lasers sources for [2-γ, 2-λ]-PDT, the advantages over [2-γ]-PDT are (i) that standard photosensitizers can be used as long as they have appropriate excited state lifetimes, and (ii) that it generates new photochemical pathways and, in particular, states that are not oxygen-dependent. This would be particularly useful in tissues that are poorly oxygenated, as are many tumours. This has been demonstrated in cell cultures.

17.5.2 Up-converting Nanoparticles

These nanoparticles have been proposed to enable efficient activation of photosensitizers while using NIR light that has greater tissue penetration than visible light. They are based typically on materials such as lanthanide- or actinide-doped transition metals that, through a 2-step charge-transfer process, sequentially absorb NIR photons and emit visible light. These nanoparticles are then administered with or conjugated to photosensitizers that are activated by visible photons. Different combinations of up-converting nanoparticles (UCNP) and photosensitizers have been reported in preclinical studies, both *in vitro* and *in vivo.* It is not clear if there is a particular advantage compared with photosensitizers that are intrinsically activatable in the NIR. Note that UNCPs represent one of several emerging applications of different forms of nanoparticles for PDT, either as energy transducers or for targeted photosensitizer delivery.

17.5.3 Bioluminescence Activation

The phenomenon of bioluminescence is common in nature, being seen in organisms ranging from fungi to insects to jellyfish. It is based on a light-emitting chemical interaction (chemiluminescence), for example, that between luciferase (*luc*) and luciferin in the presence of oxygen. Bioluminescence was introduced in the 1990s as a novel method of bioimaging, in which cells can be made luminescence by incorporating the gene that enables *luc* synthesis into the target (e.g. tumour) cells. Upon supplying luciferin, the cells, or tumours grown from them, can then be imaged by detecting the bioluminescence with a single-photon-counting detector. The light intensity is very weak, albeit in the visible range, so it was not expected that bioluminescence would be a viable light source for PDT. The first report—PDT killing of tumour cells—was in the year 2000, and the efficacy was surprisingly high. The explanation lies in the phenomenon of resonant energy transfer in which, rather than the chemical reaction generating light photons that are subsequently absorbed by the photosensitizer, activation is by direct energy transfer from the intermediate excited-*luc* state, an effect termed as BRET (bioluminescent resonant energy transfer)-PDT. The huge advantage, if this could somehow be translated into a technique in patients (for which there are many technical challenges and safety issues), is that there is no requirement for an external light source, and so, in principle, there is no limitation on the depth or volume of tissue that can be treated.

17.5.4 Ultrasound Activation

The same advantage has motivated two other novel approaches to "deep PDT", namely ultrasound activation and activation by applying ionizing radiation, both of which generate light locally within the target tissue using energy sources that have deep tissue penetration. There have been many reports in solution, cells and animal models of the so-called Sonodynamic Therapy. The results have been confusing, in part because of the difficulty in controlling and measuring the acoustic intensity within complex biological systems, especially *in vivo*. Nevertheless, some of the results are encouraging and at least one of the potential mechanisms is well understood, namely, sonoluminescence in which light is generated by ultrasound through the generation of microbubbles that then undergo rapid collapse, thereby generating very high transient local temperature that leads to light emission. (There may also be other "non-optical" activation mechanisms such as direct generation of reactive molecules from the ultrasound-photosensitizer interaction.)

17.5.5 Ionizing Radiation Activation

Finally, recent work has shown at least proof-of-concept of PDT in which the photosensitizer is activated by light generated within the tissue upon exposure to ionizing radiation. Two different approaches are possible. The first exploits the Cerenkov light generated by the passage of high-energy (megavoltage) external X-ray beams or internal radionuclides: the high-energy secondary electrons that are generated travel faster than the phase velocity of light in the tissue. (Cerenkov light is seen in the well-known blue glow in nuclear reactors). The second involves the use of scintillation nanocrystals that emit light upon absorbing X-ray energy: these are made of high atomic number materials (e.g. CeF_3) to have high X-ray cross section. Much more development is required to determine if and how either method can be adapted to generate sufficient PDT effect while using an acceptably low dose of the X-rays or radionuclides.

17.6 PDT for Cancer Treatment

From the clinical perspective, PDT is a way of producing localized tissue destruction with light after prior administration of a photosensitising agent, often requiring also the presence of free oxygen. It attracted considerable interest initially, since many photosensitising agents are retained slightly more in cancers than in the normal tissues, which raised the possibility of truly selective destruction. However, this difference is often not enough to produce truly selective tumour destruction when both tumour and normal tissue are exposed to the same light dose. Hence, there is considerable active research into ways to obtain greater tumour selectivity, but much of the spatial confinement of the PDT effect comes from how the light distribution is controlled, as discussed above. The major clinical advantage of PDT is that the biological effects are different from and complementary to other cancer treatments such as radiation therapy or chemotherapy.

Clinical studies of PDT using the photosensitizer hematoporphyrin derivative started in the late 1970s. Unfortunately, some patients were treated without an adequate understanding of the photophysics and photobiology involved so instances of overtreatment and damage to normal tissues were encountered. Skin photosensitivity was also a significant problem with earlier photosensitizers but is much less so with current agents. Minimal tissue heating occurs in PDT, due to the relatively low laser powers used ($\sim 10^2$ mW), so that connective tissues, such as collagen and elastin, are largely unaffected and the mechanical integrity of the tissue can be maintained, even if the cells are killed. This is particularly important in treating tumours in hollow organs where perforation of the wall may be dangerous. There is also evidence that PDT can spare nerves. Further, the tumour cells do not become resistant to PDT because DNA is not affected, so that multiple treatments can be given.

Ideally, optimal treatment requires "tailoring" of the light delivery, dosimetry and monitoring of both the immediate and delayed treatment responses to the individual patient and tumour. Very simple methods may work for superficial skin tumours but are often inadequate for tumours in hollow organs and for curative treatment of tumours in solid organs, because of the need to ensure complete coverage while causing minimum damage to the adjacent normal tissues. There technical issues will be considered briefly by reviewing each of the main classes of PDT treatment: superficial (e.g. skin) tumours, tumours in hollow organs and tumours in solid organs.

Skin tumours. One of the PDT compounds most widely used for skin cancers is aminolevulinic acid (ALA): this is not itself photoactive but leads to the synthesis of protoporphyrin IX (PpIX) which is both fluorescent (for imaging) and photodynamically active. ALA can be applied topically to superficial skin tumours, producing high cure rates and excellent cosmetic results. PDT with current photosensitizers is not suitable for melanomas, the most deadly form of skin cancer, largely because the melanin pigment absorbs red light too strongly. The use of 2-photon PDT (see above), together with the use of optical clearing agents that are applied to the tumour to achieve refractive index matching to reduce the optical scattering, is being explored to overcome this limitation.

Tumours of hollow organs. The relative sparing of connective tissue makes PDT suitable for treating lesions of hollow organs that have not spread beyond the wall. This is of particular value in patients who are not fit for surgery or in whom the cosmetic or functional impairment from other treatments is unacceptable. Normal bone is relatively resistant to PDT, so it may prove feasible to treat oral cancers that invade the jaw without major damage. Most localized cancers of the gastrointestinal tract and airways can be treated with surgery but PDT is an attractive alternative for patients who are unfit for surgery, and the preservation of collagen means that it is safe to produce a full-thickness PDT effect with little risk of perforation. For tumours or pre-cancers that do not invade beyond the surface mucosal layer, very high selectivity can be achieved using ALA-PpPIX PDT. In the bladder, ALA can be instilled directly through a catheter and its fluorescence enables visualization of the lesions through a cystoscope for targeting of the treatment light.

Tumours of Solid Organs. There is increasing interest in using PDT in tumours within solid organs. For example, malignant brain tumours have a much higher uptake of many photosensitizers than normal brain tissue, due to breakdown of the blood–brain barrier that prevents "foreign" materials from entering the normal brain. Hence, PDT has been used to destroy tumour remaining after surgery, as in Figure 17.9a. There has also been interest in using interstitial treatment of inoperable metastatic tumours in the brain, with fibre optics delivering laser light through small holes drilled in the skull. Sparing of connective tissues has high relevance to cancers in the lung, prostate, pancreas and head and neck. Small tumours of the major airways can be treated with PDT delivered using a flexible endoscope, with an optical fibre placed through the instrument channel (this is normally used for tissue biopsy). Tumours in the outer parts of the lungs might also be treatable by interstitial PDT, inserting the fibres either via a thin endoscope or directly through the chest wall. A number of trials have been reported for prostate and pancreas tumours, again using laser light delivered through multi-mode optical fibres, as in Figure 17.8a, under ultrasound or magnetic resonance imaging (MRI) guidance. Both the whole prostate gland and, for earlier-stage cancer, the main region of malignant tissue have been targeted. The latter has demonstrated significant advantages compared to either no treatment or radical treatments (surgery and radiotherapy) that can cause significant complications. Cancer of the pancreas is a highly fatal disease and there is no effective treatment at this time, so that PDT is a promising option.

17.7 Other Applications of PDT

PDT has been investigated for many other diseases, exploiting the common mechanism of action. The highest impact to date, in terms of number of patients and commercial success, has been in age-related macular degeneration. In this condition, there is an abnormal growth of blood vessels in the choroid, the layer that overlies the retina: when this happens near the region of highest visual acuity (the macula) it results in progressive vision loss. PDT can slow down or even partially reverse this loss. Diode laser light is delivered through a fundus camera to the involved area on the retina within a few minutes of administering the photosensitizer so that it is mainly still in the abnormal blood vessels.

Another area of great promise is treatment of localized infections, including bacterial infection. In this case cationic photosensitizers are used that have some selectivity for uptake into the bacterial wall. Treatment of chronic infection of the gums is in use in some countries, with both photosensitizer and light delivered into the "pockets" between the teeth and the affected gum. Treatment of chronic-infected skin wounds, such as that often occur on the legs in advanced diabetes, has also been investigated in clinical trials and shows promise. In this case, either a diode laser or LED arrays can be used as the light source. PDT has been used to remove contamination in blood transfusion, where several orders-of-magnitude of viral inactivation can be achieved with minimal damage to normal blood cells. In this case, special irradiators have been developed to apply the light to the whole blood bag. A similar approach is being investigated for selective viral and bacterial contamination in organs for transplantation, to reduce the risk of passing infection from the donor to the recipient. PDT has also shown high efficacy in treating fungal infections. A major advantage of PDT in infectious diseases is that it has minimal induction of resistance that limits many antibiotic drugs. This is because it kills the microorganisms by damaging the cell walls without causing genetic damage, so that new mutations or selection of pre-existing mutations that provide resistance are not induced.

17.8 Conclusions

PDT offers many challenges and opportunities for existing and new optical technologies, including but not limited to lasers. To date, most research and clinical applications have been based on CW light sources and particularly diode lasers in the visible/NIR range with powers of typically several Watts. However, in the future it is likely that pulsed laser sources will also be required, particularly for 2-photon photosensitizer activation. Fibre-optic-based devices for light delivery are critical for many applications, and there is an increasing need to tailor these to specific body sites. Scanning laser beam technologies may also be of value for light delivery in some special cases, such as treatment of retinal diseases, and may also be useful for *in vivo* imaging for dosimetry and monitoring of treatment response. State-of-the-art optical technologies will play an increasingly important role in photophysical and biological-response measurements of the PDT processes in cells and tissue. Finally, an aspect mentioned briefly above is the use of the fluorescence characteristics of many photosensitizers. Fluorescence spectroscopy or imaging can be used either for

PDT dosimetry, as in determining the concentration of drug in the tissues or the bleaching of the drug during irradiation, for target tissue localization or, more generally, for disease detection and therapeutic guidance.

In the 1980s, the clinical applications of PDT, e.g. in tumour therapy, were crude and depended on large, expensive, unreliable and difficult-to-use lasers such as argon ion-pumped dye lasers and excimer lasers. Many clinical lasers are now the size of a briefcase, can be plugged into a standard power socket and are rugged enough to be moved easily between different sites. The applications are becoming more sophisticated as the biological effects are better understood, and the basic principle of PDT can be applied to a wide range of disease processes.

Bibliography

The following articles are starting points into the literature on the biophysics and optical instrumentation of PDT. In addition, some articles give an overview of the clinical applications and/or photobiology of PDT.

Ackroyd R, Kelty C, Brown N and Reed M. The history of photodetection and photodynamic therapy. *Photochemistry and Photobiology* 74, 656–669, 2001.

Presents an historical perspective on the development of photoactivated drugs, both for therapy and as diagnostic agents.

Agostinis P et al., Photodynamic therapy of cancer—An update. *CA Cancer Journal for Clinicians* 61, 250–281, 2011.

A review of the roles and status of PDT in cancer treatment.

Brahmavar S, Chen Q, Zhu TC, Hetzel FW, Jacques SL, Patterson MS and Wilson BC, Photodynamic Therapy Dosimetry, American Association of Physicists in Medicine, Report #88, 2005.

A guide to practical dosimetry for clinical PDT.

Brancaleon L and Moseley H. Laser and non-laser source for photodynamic therapy. *Lasers in Medical Science* 17, 173–186, 2002.

A critical summary of light sources for PDT.

Chitgupi U, Qin Y and Lovell JF. Targeted nanomaterials for phototherapy. *Nanotheranostics.* 1, 38–58, 2017.

A review of nanoparticles for use in light-based treatments including PDT.

Cochrane C, Mordon SR, Lesage JC and Koncar V. New design of textile light diffusers for photodynamic therapy. *Materials Science and Engineering C—Materials for Biological Applications* 33, 1170–1175, 2013.

Describes a large-area light diffusing device for PDT.

Fisher WG, Partridge WP Jr, Dees C and Wachter EA. Simultaneous two-photon activation of type-I photodynamic therapy agents. *Photochemistry and Photobiology* 66, 141–55, 1997.

Discusses 2-photon PDT, both in terms of the photophysics/photochemistry and the light source requirements.

Gemmell NR, McCarthy A, Kim M, Veilleux I, Zhu T, Buller GS, Wilson BC and Hadfield RH, A compact fiber optic probe-based singlet oxygen luminescence detection system. *Journal of Biophotonics* 10, 320–326, 2017.

Describes the development and use of superconducting nanowire singlet-oxygen detectors.

Karotki A, Khurana M, Lepock JR and Wilson BC, Simultaneous two-photon excitation of photofrin in relation to photodynamic therapy. *Photochemistry and Photobiology* 82, 443–452, 2006.

A study of simultaneous 2-photon PDT photophysics.

Swartling J, Axelsson J, Ahlgren G, Kälkner KM, Nilsson S, Svanberg S, Svanberg K, Andersson-Engels S. System for interstitial photodynamic therapy for online dosimetry: first clinical experiences with prostate cancer. *Journal of Biomedical Optics* 15, 058003, 2010.

A novel instrument for PDT, based on optical feedback to control light distribution to multiple fibres.

Oleinick NL and Evans HH. The photobiology or photodynamic therapy—cellular targets and mechanisms. *Radiation Research* 150, S146–156, 1998.

A comprehensive overview of the underlying cellular and molecular mechanisms of PDT.

Smith G, McGimpsey WG, Lynch MC, Kochevar IE and Redmond RW. An efficient oxygen independent two photon photosensitization mechanism. *Photochemistry and Photobiology* 59, 135–139, 1994.

A demonstration of 2-photon/2-color PDT.

Wilson BC and Patterson MS, The physics, biophysics and technology of photodynamic therapy. *Physics in Medicine and Biology* 53, R61–109, 2008.

A detailed review of the biophysical and technological aspects of PDT.

18

Therapeutic Applications: Thermal Treatment of Tumours

Stephen G. Bown

CONTENTS

18.1 Introduction 261
18.2 Laser Therapy with Flexible Endoscopes 262
    18.2.1 Cancers of the Gastrointestinal Tract 262
    18.2.2 Cancer of the Lungs 263
    18.2.3 Urology 263
18.3 Interstitial Laser Photocoagulation 263
18.4 Conclusion 265
References 265

18.1 Introduction

The most important effects of laser light used in treating tumours (benign or malignant) are thermal and photochemical. The thermal applications are the simplest to understand; the effect depending on how much heat is delivered, how fast it is delivered, and the volume of tissue in which it is absorbed. They are discussed in this chapter.

For general surgeons, their first hope was that lasers could produce a clean cut in living tissue and prevent bleeding from the edges of the incision. It soon became clear that one could not have one's cake and eat it. To stop bleeding, the heat from the laser needed to coagulate the tissue in the incision edge: the larger the blood vessels in the tissue were, the deeper the coagulation needed to go. The first laser to be used in this way was the carbon dioxide laser (10 600 nm) with a beam which is strongly absorbed in water, so only penetrates a fraction of a mm into tissue. Consequently, it can cut or vaporize tissue with little damage to underlying layers, but it can only stop bleeding from blood vessels below about 0.2 mm in diameter and the beam cannot be transmitted via flexible fibres that are suitable for clinical use. This laser is well established as a non-contact knife in relatively inaccessible areas like the brain, spiral cord, mouth, and upper airways (including the pharynx and larynx) and for ablating small lesions on sites like the skin or cervix (neck of the womb) but its use is limited in other areas, as the beam can only be delivered to sites accessible in a straight line from the outside [1].

Near-infrared light, as from a Nd:YAG (neodymium yttrium aluminium garnet) or a semiconductor diode laser, penetrates tissue much better, producing effects through up to 10 mm of tissue, which is much better for stopping bleeding in soft tissue as it seals small blood vessels up to about 1 mm in diameter by thermal contraction of the surrounding tissue. Light in this wavelength region is much more easily transmitted via flexible fibres. The effects depend on the laser power, target area, and exposure time. The range of changes seen is shown in Table 18.1. The key ways these are used in clinical practice are shown in Table 18.2, and the most commonly used types of thermal laser are shown in Table 18.3.

TABLE 18.1

Range of Thermal Effects on Biological Tissue

Total Destruction:	Immediate Vaporization
	Cell death with later sloughing
Destruction with reconstruction:	Cell death with healing by scarring
	Cell death with healing by regeneration
Reversible:	Swelling and inflammation
	Local warming only

TABLE 18.2

Thermal Laser Effects Used in Tumour Therapy

Heat at high power:	Debulking of tissue (vaporization and coagulation) and non-contact cutting
Heat at low power: (ILP)	Gentle coagulation of lesions contained within solid organs

TABLE 18.3

Thermal Lasers Used Most Frequently in Tumour Therapy

	Wavelength (nm)	Fibre Transmission	Properties
Carbon dioxide (CO_2)	10 600	No	Precise, no contact
(see chanter B3.1)		Incision, relatively	Poor control of bleeding
Neodymium YAG (Nd:YAG)	1064	Yes	Vaporization, deep
		Coagulation, good	Control of bleeding
Diode	630–900	Yes	Simple reliable
		Similar histological effect to Nd:YAG,	
			But lower power

18.2 Laser Therapy with Flexible Endoscopes

Endoscopes are flexible telescopes that can be used to examine many of the hollow organs of the body such as the lungs, stomach and bladder, accessed via natural orifices. Previously, haemorrhage from peptic ulcers used to be a common medical emergency and the specialists looking after such patients, gastroenterologists, were searching for an endoscopic technique to stop such haemorrhage without the need for open surgery. They showed that endoscopic treatment with the Nd:YAG laser could, indeed, significantly reduce the need for surgery. It is of interest historically that the 'glamour' of this early application of lasers led to a much better understanding of what made ulcers bleed. The key discovery was that if endoscopic examination of an ulcer that had just bled revealed an artery in the ulcer crater, there was a much higher risk of it continuing to bleed compared with the one in which no artery was seen. In the latter case, bleeding almost always stopped spontaneously. The artery also provided a target for endoscopic laser therapy. More recently, a simpler, cheaper and less glamorous technique has been shown to be just as effective. This technique, sclerotherapy, involves the endoscopic injection of adrenaline around the artery that has bled in the ulcer crater. Further, better understanding of the cause of peptic ulcers associated with the development of more effective drugs has drastically reduced the incidence of this problem. The best established endoscopic application of this laser now is to relieve the worst symptoms of advanced cancers that are blocking the oesophagus, rectum or major airways.

18.2.1 Cancers of the Gastrointestinal Tract

Most patients with cancers of the oesophagus (gullet) and the gastric cardia (the region where the oesophagus joins the stomach) have no prospect of cure at the time they first get symptoms due to the extent of their disease. The dominant symptom is usually an inability to swallow (dysphagia). Typically, this starts as a difficulty with solids and then progresses to problems with liquids. The therapeutic challenge is to relieve this problem with the minimum of general upset to these unfortunate patients who, in most cases, have less than a year to live.

Endoscopic Nd:YAG laser recanalization of advanced oesophageal cancers was first reported in 1982. Only patients with obvious lumps of tumour visible down the endoscope are suitable. The aim is to reduce the bulk of these lumps, concentrating on those parts causing the worst obstruction. Prominent nodules can be vaporized. Smaller nodules can be coagulated and the dead tissue sloughs over a few days. The deep penetration of the Nd:YAG laser beam prevents bleeding from the underlying tissues but care must be taken not to go too deep as this could cause a perforation [2].

The main problem is the need to repeat laser treatment every 4–6 weeks to maintain good palliation. However, data are now available to show that if laser therapy is followed by radiotherapy, with or without chemotherapy, the period of good palliation can be considerably prolonged. One of the simplest options is brachytherapy, in which a radioactive wire is positioned in the cancer for a single treatment of less than half an hour. This can extend the period of palliation from 5 weeks to 5 months, although it does not increase the survival time. The other main technique for palliating these advanced cancers is to put a short tube in to splint the cancer open. With tubes or laser or a combination, most of these patients can be kept swallowing reasonably for most of their remaining life.

Most cancers of the rectum and lower bowel are best treated surgically but 5%–10% are unsuitable for resection due to the general condition of the patient or the extent of malignant disease. Diverting the faecal stream with a colostomy (bringing the bowel out to a bag on the abdominal wall to bypass the cancer) palliates obstruction but does not relieve bleeding or discharge of mucus (slime) from advanced rectal cancers. Endoscopic laser therapy is a simple, out-patient way of helping these patients and can often keep them comfortable for several months or even a year or two [3].

18.2.2 Cancer of the Lungs

Endoscopic laser palliation of advanced lung cancers has developed in parallel to applications in the gastrointestinal tract. As many as 50% of patients with inoperable lung cancer cough up blood at some stage of their illness and about 60% become short of breath from tumour obstruction of a major airway. For patients with symptoms due primarily to endoscopically accessible lumps of tumour in the major airways, Nd:YAG laser therapy can play a useful palliative role.

The procedure is technically more difficult and potentially more hazardous than in the gastrointestinal tract. Temporary total occlusion of the oesophagus is tolerable but of the trachea is not! Further, the gastrointestinal tract has a convenient route for disposing of dead debris, which the airways do not. Procedures in the gastrointestinal tract are usually performed under sedation but lung treatments require general anaesthesia. This has proved remarkably safe despite the frequent coexistence of other major smoking-related diseases. Insertion of a rigid bronchoscope (basically, a hollow metal tube that goes down to where the trachea divides into the main airways for each lung) gives access for ventilating the patient, sucking out blood and secretions and clearing debris. Laser therapy is undertaken with a flexible endoscope inserted through the rigid one. Coagulated tissue is removed immediately with large forceps.

Response to therapy is immediate and can be dramatic. Best results are achieved with tumours in the largest airways with viable lung beyond the obstruction (Figure 18.1). Occasionally, resuscitative treatment of an obstructing tracheal tumour can pave the way for potentially curative surgery [4]. As with oesophageal cancers, radiotherapy can prolong the period of good palliation after laser recanalization of cancers of the major airways.

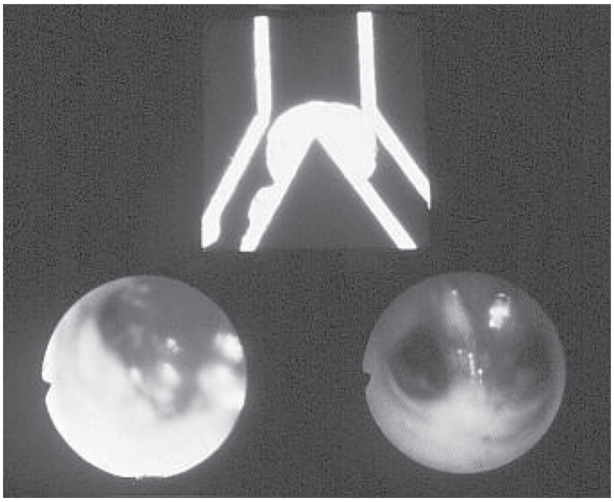


FIGURE 18.1 Endoscopic treatment of an advanced lung cancer occurring at the site most vulnerable to tobacco toxins, where the main airway, the trachea, divides into the airway going into each lung. At the top is a diagram showing how the tumour is completely blocking the airway to one lung and partially blocking the airway to the other lung. The lower photograph on the left shows the view down the endoscope, with the patient surviving by breathing through a small airway into one lung. After laser therapy, a good airway has been restored to both lungs, as seen in the photograph on the right. This treatment was not curative but provided temporary relief of the patient's most distressing symptom of breathlessness. (Figure courtesy of Dr Martin Hetzel.)

18.2.3 Urology

The carbon dioxide and Nd:YAG lasers have been used to treat lesions of the external genitalia such as warts and superficial penile cancer. Endoscopically, the main target is bladder tumours. Nd:YAG and near-infrared diode lasers may have a role in the management of benign enlargement of the prostate but conventional surgical techniques work just as well.

Unlike cancers of the oesophagus and lungs, bladder tumours usually cause symptoms at a much earlier stage of their development, as they bleed into urine. These can be treated endoscopically either by conventional electrocoagulation or with a laser. The greater control and precision of laser treatment makes it an attractive option for ablation of small tumours under local anaesthesia, although it is not widely used.

18.3 Interstitial Laser Photocoagulation

The endoscopic applications of the Nd:YAG laser described earlier use high powers (up to about 70 W) and, although effective, are relatively crude. The concept of interstitial laser photocoagulation (ILP) is quite different. One or more laser fibres are inserted through needles directly into a lesion in the centre of a solid organ. The laser is then activated to deliver a low power beam, typically 2–3 W per fibre for 300–500 s. This causes gentle thermal coagulation without vaporization, centred around each fibre tip, with no more damage to the overlying normal tissues than the minimal trauma of inserting the needles. The technique was first described in 1983 and might be described as the 'slow casserole' approach, rather than the 'stir fry' of the high-power applications. Subsequently, a series of experiments showed that it was possible to produce an area of tissue destruction about 1.5 cm in diameter around each fibre site in normal liver and that these lesions healed safely, mainly by regeneration of the normal liver. This led to the first clinical report of the percutaneous (i.e. using laser fibres through needles inserted through the skin) treatment of small tumours of the liver in 1989. Larger lesions require more fibres as the volume of the tissue destruction that can be produced around each fibre is limited (Figure 18.2).

As the aim of ILP is to treat lesions within solid organs, it is essential that good imaging is available to position fibres correctly and to assess the results both at the time of treatment and during follow-up. It must be possible to define the limits of the target lesion and to match the treatment to these limits to maximize the prospects of destroying the entire lesion and minimize the risk of unacceptable damage to adjacent normal tissues. Ultrasound scanning has proved valuable for inserting fibres but is poor at documenting the tissue response. Contrast-enhanced CT (computerized tomography) scans are useful but only give a true picture of the extent of laser-induced tissue destruction 24 h or longer after treatment. Laser-induced thermal changes have been monitored in real time with dynamic magnetic resonance imaging (MRI), as optical fibres do not

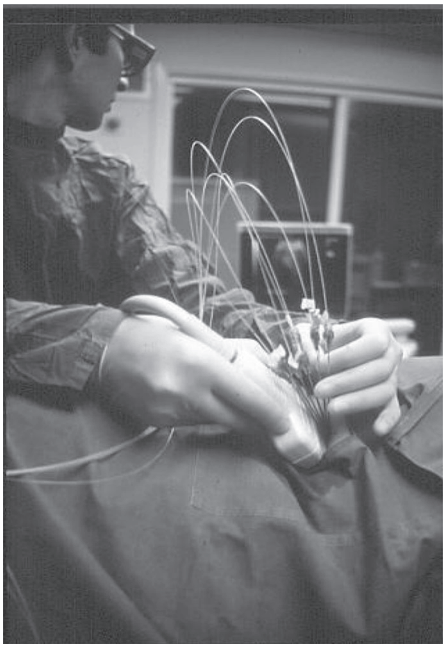


FIGURE 18.2 ILP of a tumour in the liver. Under guidance from an ultrasound scanner (the ultrasound probe being held by the doctor), a series of needles are inserted through the skin into the tumour within the liver. Laser fibres are then inserted through the needles for direct coagulation of the tumour.

cause artefacts on MRI scans [5]. The MRI indicates temperature changes in the area being treated, which can be correlated with the final extent of laser-induced tissue destruction. Nevertheless, this technique is still at an early stage of development and is a long way from being available for routine use.

An increasing number of therapeutic options are being considered for the localized destruction of the diseased tissue within solid organs. In addition to ILP, these include direct injection of absolute alcohol or anti-cancer drugs, heating by microwaves, radiofrequency waves or high-intensity focused ultrasound, cryotherapy (freezing down to the temperature of liquid nitrogen), brachytherapy (the insertion of strongly radioactive sources such as an Ir^{192} wire for a short period of time, typically 20–30 min) or photodynamic therapy (PDT). ILP and PDT have the advantage that the fibres required for light delivery are thin and can be inserted percutaneously through small needles, so they can be positioned relatively easily within firm tissue, such as in secondary cancers. Further, the area of necrosis produced is well defined, roughly ellipsoidal and does not depend on the texture of the tissue as much as with the other techniques. Thermal laser effects and photodynamic effects are compared in Table 18.4. Radiofrequency heating may give comparable results but the necrosis produced is more dependent on the electrical properties of the target tissue. It is difficult to inject alcohol into hard tissue, although this is a much cheaper option and works reasonably well for softer lesions like primary liver cancers. Focused ultrasound uses phased arrays of transducers, approaching the target zone from multiple directions and is totally non-invasive but it is difficult to use on deep tissues like the liver and is more appropriate for superficial tissues like the breast.

ILP could have any role in the management of small, localized breast cancers, as an alternative to lumpectomy (the surgical removal of just the part of the breast containing the cancer, without removing the entire breast). The attractions would be that treatment would leave no scar or change in the shape of the breast and could be carried out just with sedation and local anaesthetic under image guidance as a day case procedure. Studies treating small breast cancers with ILP prior to conventional surgical excision have shown that contrast-enhanced MRI scanning can define both the tumour limits and the limits of laser-induced tissue destruction remarkably accurately [6]. However, a lot more work is required to clarify whether ILP could be used as an alternative to lumpectomy. This really centres on establishing whether the imaging is good enough to pick up any areas of cancer that the ILP may have missed. It should be remembered that local removal of the cancer is only part of the treatment for such patients and many will also require radiotherapy to the breast and sometimes anti-cancer drugs as well.

ILP has the potential to treat lesions within any solid organ that are not suitable for surgical excision and which are accessible for image-guided needle insertion. There are isolated reports of its application to cancers in the outer parts of the lungs (i.e. areas beyond the reach of a flexible endoscope passed down the airways) in patients who are not fit for surgery and to lesions within the brain. It is becoming a routine treatment for fibroadenomas of the breast (benign lumps that

TABLE 18.4

Comparison of ILP and Photodynamic Therapy (PDT)

	ILP	PDT
Nature of biological effect	Thermal	Photochemical
Effect on connective tissue	Destroyed	Mainly unaffected
Healing	Resorption and scarring, some regeneration	Regeneration, sometimes with scarring (particularly in muscle)
Selectivity of effect between tumour and tissue of origin of tumour	None	Minimal
Selectivity of effect between tumour and other adjacent tissues	None	Possible between mucosa and underlying muscle in hollow organs
Wavelength of light used	Infrared (805–1064 nm)	Red (630–760 nm)
Typical laser power per fibre	3–5 W	0.1–0.3 W (higher for illuminating hollow organs)

are common in young women) and for benign bone tumours (osteoid osteomas). It is under assessment for treating fibroids of the uterus (common benign growths that can cause heavy and painful periods and may lead to infertility) and has been used for benign prostatic hypertrophy (enlargement of the prostate that may cause urinary difficulties in older men). In principle, ILP could be used to treat any tumour of a solid organ where the limits of the lesion can be accurately defined and the surrounding normal tissue can tolerate a small amount of localized thermal damage. ILP is a technique still in the evolutionary stage, and there are few large studies with long-term follow-up but it does appear safe, effective and relatively simple for appropriate conditions. It also has the advantage of being repeatable.

18.4 Conclusion

The basic principles of thermal laser therapy can be applied to disease processes in a wide range of organs. However, it must be remembered that all laser treatments are local treatments. The challenge is always to define the true extent of the disease process and then match the extent of the treatment to this so that the disease process is controlled and all treated tissue maintains adequate structure and function during healing. Laser therapy is not suitable for cancers that have spread to many different sites around the body.

There are increasing numbers of tumours (benign and malignant) for which thermal laser therapy can offer a minimally invasive therapeutic option that is often more comfortable for the patient with a more rapid recovery time than conventional alternatives such as open surgery. For some benign tumours, such as fibroadenomas of the breast, laser therapy on its own may be all that is required. However, for cancers, in most cases, thermal laser therapy will only be part of the overall treatment plan for individual patients and will be complementary rather than competitive to current options, such as chemotherapy (anti-cancer drugs) and radiotherapy. This is particularly true for situations like the palliation of swallowing difficulties with oesophageal cancers, when laser and brachytherapy together can offer better palliation than either on its own.

REFERENCES

1. Ambrosch P, Kron M and Steiner W 1998 Carbon dioxide laser microsurgery for early supraglottic carcinoma *Ann. Otol. Rhinol. Laryngol.* **107** 680–8.
2. Ell C, Riemann J F, Lux G and Demling L 1986 Palliative laser treatment of malignant stenoses in the upper gastrointestinal tract *Endoscopy* **18** 21–6.
3. Brunetaud J M, Manoury V, Ducrotte P, Cochelard D, Cortot A and Paris J C 1987 Palliative treatment of rectosigmoid carcinomas by laser endoscopic photoablation *Gastroenterology* **92** 663–8.
4. Hetzel M R and Smith S G T 1991 Endoscopic palliation of tracheobronchial malignancies *Thorax* **46** 325–33.
5. Vogl T J, Mack M G, Straub R, Roggan A and Felix R 1997 Magnetic resonance imaging guided abdominal interventional radiology: laser induced thermotherapy of liver metastases *Endoscopy* **29** 577–83.
6. Mumtaz H, Hall-Craggs M A, Wotherspoon A, Paley M, Buonaccorsi G, Amin Z, Wilkinson I, Kissin M W, Davidson T I, Taylor I and Bown S G 1996 Laser therapy for breast cancer: MR imaging and histopathologic correlation *Radiology* **200** 651–8.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

19

Therapeutic Applications: Dermatology—Selective Photothermolysis

Sean Lanigan

CONTENTS

19.1 Introduction 267
  19.1.1 Photothermal Interactions 267
  19.1.2 Selective Photothermolysis 268
19.2 Laser Treatment of Cutaneous Vascular Lesions 268
  19.2.1 Principles of Selective Photothermolysis and the Treatment of PWSs 268
19.3 Laser Treatment of Pigmented Lesions and Tattoos 270
19.4 Laser Treatment of Hair by Selective Photothermolysis 270
19.5 Carbon Dioxide and Erbium:YAG Lasers in Dermatology 271
  19.5.1 Erbium:YAG Laser in Dermatology 271
References 272
Further Reading 272

19.1 Introduction

When laser light interacts with the skin it can be reflected, scattered, transmitted or absorbed. Laser–skin interactions after light absorption can be grouped as photothermal, photochemical, plasma-induced ablation, photoablation and photodisruption.

By far the most important interaction in dermatology is photothermal. Photochemical interactions are important in photodynamic therapy. Experimentally, excimer lasers produce some of their effects by photoablation.

19.1.1 Photothermal Interactions

In photothermal laser interactions the absorbed photons are converted to heat. This local increase in temperature is the most significant influencing factor. The effect of heat on biological tissue depends on the duration and the peak value of the tissue temperature achieved. The effects of heating can be seen as *coagulation*, which can proceed to *necrosis* and *vaporization*, resulting in tissue *ablation* and *carbonization* (Figure 19.1).

As tissue is heated, structural changes occur in complex molecules such as proteins, DNA and RNA. These structural changes result in an impairment of function termed *denaturation*. In addition to denaturation, there is grosser structural disorder with the entanglement of molecules, termed coagulation. Denaturation and coagulation generally proceed as tissue temperature rises above 60°. Again the duration of temperature rise is also important in inducing tissue damage. Thermal coagulation causes cell necrosis. Gross thermal necrosis of the dermis is equivalent to a burn, and in many instances, laser surgery involves controlled and limited thermal injury to the skin.

Heating tissue above 100° causes evaporation of water and vaporization of tissue. Tissue vaporization is used therapeutically

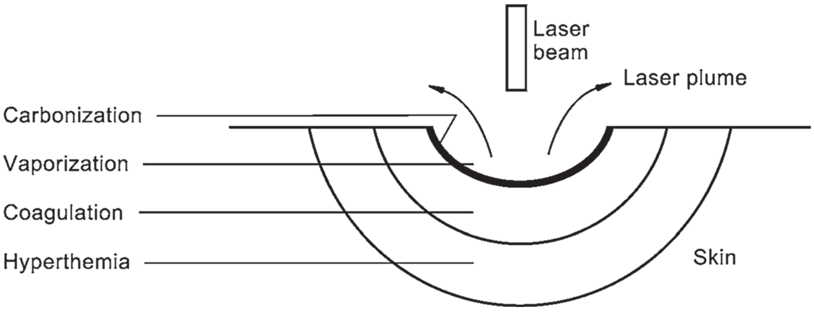


FIGURE 19.1 Photothermal laser effects. The width of each zone varies with the laser parameters.

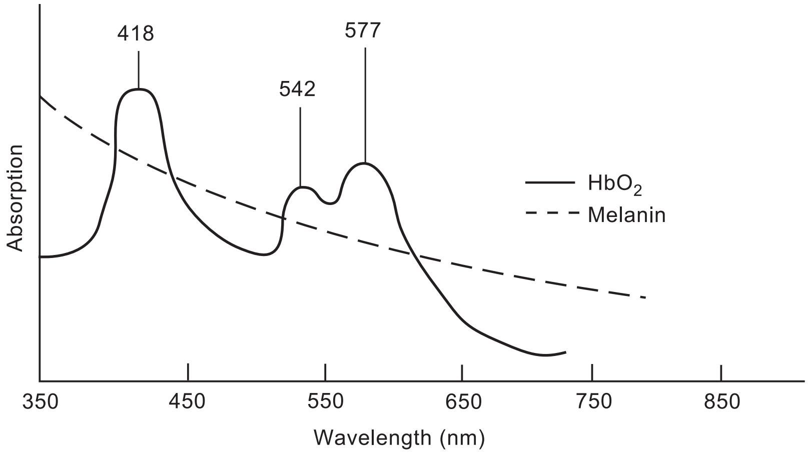


FIGURE 19.2 Schematic diagram of absorption of oxyhaemoglobin (HbO_2) and melanin.

in resurfacing lasers such as the CO_2 laser. Vaporized tissue is seen as a plume of steam arising from the treated area containing water vapour and particulate material. If the energy from the laser is delivered in a sufficiently short duration, all the absorbed energies will be utilized in vaporizing the tissue water in the superficial layer of irradiated tissue. The treated layer is rapidly vaporized away so that there is no heat source applied to the deeper tissue layers. In this way, by using short-pulsed CO_2 or Er:YAG lasers, the residual thermal damage below the vaporized tissue is only a few micrometres in depth.

In contrast, if the pulse durations employed for vaporization are long, there is heat conduction to deeper tissues. The surface layer becomes desiccated and extreme heat rises occur in excess of 350° with charring (carbonization) of tissue. The tissue bed below a long-pulse (continuous wave) CO_2 laser can be thermally necrosed to a depth of 1 mm or more. This is likely to result in visible scarring of healed tissue. In general, carbonization is an unwanted outcome of laser treatment and usually arises from inappropriate laser parameter selection or faulty technique.

19.1.2 Selective Photothermolysis

Perhaps the most important concepts in the development of lasers in dermatology have been elaborated by Anderson and Parrish, in their paper published in *Science* (1983) [1]. In this paper the authors discussed the concept of *selective photothermolysis*. Utilizing the principles of selective photothermolysis by the appropriate selection of wavelength, pulse duration and energy fluence, a precise biological target can be thermally damaged by lasers.

In considering the most appropriate wavelength for the target, the light must penetrate sufficiently to reach the target and also be absorbed by it to produce photothermal changes. For example, oxyhaemoglobin absorbs visible light with absorption peaks at 418, 542 and 577 nm. However, for therapeutic thermal damage of cutaneous blood vessels, the more deeply penetrating wavelengths are preferred.

The heat generated within the target can be confined to that target by the appropriate selection of pulse duration. This is related to the thermal relaxation time T_r of the target.

$$T_r = d^2/4k \tag{19.1}$$

where d is the layer of thickness and k is the thermal diffusivity. The thermal relaxation time is a measure of the intrinsic cooling time of the target and is the time taken for the target to dissipate half of the incident thermal energy. This cooling time is primarily related to the physical size of the target: a larger target gives a longer thermal relaxation time. If the laser pulse duration is equal to or less than the thermal relaxation time of the target then unwanted heat diffusion to adjacent tissue will be reduced. For example, for port-wine stain (PWS) capillaries pulse durations around 1 ms would be appropriate, whereas for melanosomes in pigmented lesions nanosecond pulses would be more appropriate. Having defined the most appropriate wavelength and pulse duration, it is then important to deliver sufficient fluence to the target to produce the desired temperature rise. The fluence will be influenced by attenuation of the beam by competing chromophores, e.g. melanin in the epidermis overlying PWSs, depth of penetration and scattering.

In clinical practice, many targets have a three-dimensional structure which includes depth and it is impossible to achieve complete thermal damage with one laser impact as superficial components of the target will absorb light, shadowing deeper targets. This is why for many structures, e.g. PWSs, pigmented naevi and tattoos, multiple treatments are necessary.

19.2 Laser Treatment of Cutaneous Vascular Lesions

19.2.1 Principles of Selective Photothermolysis and the Treatment of PWSs

Current successful treatment of PWSs follows the principles of selective photothermolysis proposed by Anderson and Parrish [1,2]. The target in PWS is haemoglobin in cutaneous blood vessels. PWSs are benign, vascular naevi consisting of ectatic dermal capillaries situated predominantly superficially in the dermis. The colour of the PWS is related in part to the erythrocyte content of the ectatic capillaries [3].

Laser treatment that successfully reduces the number, size and erythrocyte content of these vessels results in observable lightening of the PWS.

The absorption spectra of oxyhaemoglobin and melanin are shown in Figure 19.2. For selective destruction of red cells, light preferentially absorbed by haemoglobin is necessary. It can be seen from the figure that haemoglobin has three major absorption peaks. Although 418 nm is the strongest absorption peak the penetration of light at this shortest wavelength is not sufficient to treat the dermal capillaries. In addition, melanin, which is a competing absorber of light in the visible range, has a falling absorption capacity with longer wavelengths. Light at 577 nm is *selectively* absorbed by haemoglobin compared to other cutaneous structures and will penetrate into dermal vessels. When laser energy of sufficient fluence is absorbed by haemoglobin, local heating will occur from within the blood vessel [4]. Depending on the duration of the heat application, this temperature rise will diffuse outwards to the vessel wall and beyond to the perivascular connective tissue. The extent of heat diffusion can be estimated with reference to the *thermal relaxation time*: in PWS, vessel diameters can range from 10 to 100μm. The thermal relaxation time of vessels in this size range has been measured as 1–10 ms [5]. If the pulse duration of the laser greatly exceeds the thermal relaxation time of the targeted blood vessel, non-selective thermal damage will occur to perivascular connective tissue and beyond, leading to adverse clinical events such as scarring. Conversely, if the pulse duration is very short, vessels can be punctured by the laser beam with haemorrhage but the duration of heating is insufficient to result in permanent damage to blood vessels with no resulting lightening of the PWS [6]. See Figure 19.3.

Other factors to consider in the selective photothermolysis of PWS are as follows:

1. *Epidermal melanin* [7]. In patients with darker skin types more of the laser energy will be absorbed within the pigmented epidermis. This can result in insufficient energy reaching the blood vessels and an increased incidence of unwanted post-inflammatory pigmentary changes. This can be overcome to some extent by dynamic cooling of the epidermis [8]. The epidermis is cooled selectively by a spurt of cryogen. This reduces epidermal temperature rise without affecting thermal injury to dermal vessels.
2. *Beam diameter* [9]. Many of the observations relating to the interactions of laser light and skin have been made from mathematical models. The effects of diffusion losses at the edges of a laser beam can contribute to loss of energy in the centre with very small beam diameters. Conversely with wide beams, backscattered irradiance augments the central beam energy. In general terms, large laser beam diameters have deeper penetration than smaller beams when the same energy fluence is delivered.
3. *Shadowing effects* [10]. As discussed previously, light penetrating the skin can be absorbed and scattered. Coherent laser light penetrates the skin from above and penetrates centrally; the light is attenuated with distance. When strongly absorbent molecules interact with the light, no further penetration occurs. In a PWS, there is a three-dimensional matrix of blood vessels. If strong absorption of light occurs in the most superficial capillaries there will be insufficient penetrating light to effect thermal damage to deeper vessels. For this reason 585 nm light has been used in tunable dye lasers instead of 577 nm. This longer wavelength light penetrates more deeply and is less well absorbed by melanin but also less avidly absorbed by haemoglobin in superficial capillaries, so some light is scattered downwards to deeper blood vessels. Tan et al. [11] demonstrated that by increasing the wavelength of the pulsed dye laser from 577 to 585 nm, the depth of penetration increased from 0.5 to 1.2 mm while still retaining selectivity of cutaneous blood vessels.

In fact, the interactions of laser light and PWS are extremely complex. Clinical and histological assessments of PWS laser interaction have demonstrated the efficacy of laser treatment, but it is acknowledged that PWS are far more heterogeneous than originally assumed.

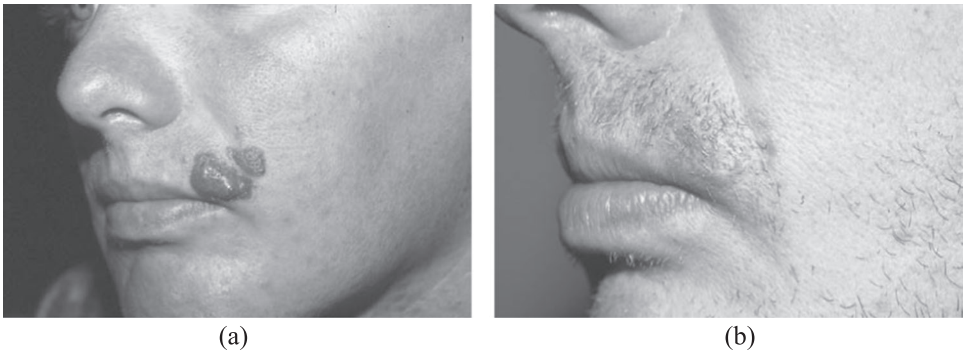

(a) (b)

FIGURE 19.3 (a) Hypertrophic port-wine stain. (b) Marked improvement after six treatments with the pulsed dye laser.

19.3 Laser Treatment of Pigmented Lesions and Tattoos

The principles of selective photothermolysis can also be applied to the laser treatment of pigmented cutaneous lesions. The target chromophore in pigmented lesions is melanin, which is packaged within cells in melanosomes. These organelles are 0.3–1.0 µm in length. Melanin absorbs light over a broad spectrum from 250 to 1200 nm so lasers emitting light within this range can be expected to affect cutaneous pigmentation to some degree. However, there are other chromophores competing for light particularly in the visible range, e.g. haemoglobin, and tissue water in the infrared. A therapeutic window of wavelengths exists between 630 and 1100 nm where melanin absorption exceeds haemoglobin absorption and the light penetrates deeply. As the absorption of melanin falls with increasing wavelength, higher fluences are required at longer wavelengths of light to effect the same damage to melanosomes as lower fluences at shorter wavelengths. The advantage of longer wavelengths of light is deeper dermal penetration.

To effect organelle-specific damage to melanosomes without thermal damage to adjacent structures, the pulse duration of the laser should be within the thermal relaxation time of the melanosome, which is estimated at 100–1000 ns. Lasers with pulse durations less than 1 μs can effect selective photothermolysis of melanosomes. Damage occurs both from thermal effects but also shock waves arising from rapid thermal-induced expansion of tissue occurring during the short-pulse durations.

Tattoos consist of intra-dermal, intra-cellular insoluble ink particles. In general, the same Q-switched lasers used for pigmented lesions are also appropriate for tattoos. Although picosecond pulsewidths may be more successful, lasers with these short pulse-widths are not commercially available [12]. Tattoo clearance after laser treatment is often incomplete and in part reflects the varying absorption of light by different tattoo inks seen in professional tattoos [13] (see Figure 19.4).

19.4 Laser Treatment of Hair by Selective Photothermolysis

To fully appreciate the relevance of laser hair interactions, the reader should be familiar with hair anatomy, hair growth cycles and the variation in hair growth cycles in different parts of the body.

The base of the hair follicle is 2–7 mm below the surface of skin so only red and near-infrared wavelengths of light will penetrate to this depth. Active hair growth occurs during *anagen*. It emerges from the skin surface and continues to grow until the onset of *catagen*. The terminal portion of the hair becomes club-shaped and lacks melanin. The club hair moves upwards as the follicle shortens and moves into *telogen*, the resting phase of the hair cycle. The follicle re-enters anagen spontaneously at the end of telogen or may be induced to do so if the hair is plucked. The new hair grows up below and then beside the club hair. In regions of the body other than the scalp anagen is relatively short and telogen is relatively long. Only hairs in anagen are susceptible to injury from lasers [14]. Length of hair is related to duration of anagen and rate of hair growth. Hair growth is fastest for longest growing hairs (beard, scalp) and slowest for shortest hairs (eyebrows, thighs).

Several lasers have been developed for selective photothermolysis of hair follicles (Table 19.1). Melanin in the hair follicle is the selected chromophore. The pulse duration appropriate for the thermal relaxation time of hair follicles ranges from several to 100 ms. The site within the hair follicle where the stem cells necessary for the regeneration of the follicle

TABLE 19.1

Lasers for Hair Removal

Laser	Wavelength (nm)	Pulse (ms)
Ruby	694	1–3
Alexandrite	755	2–40
Diode	800	5–30
Nd:YAG	1064	3–50

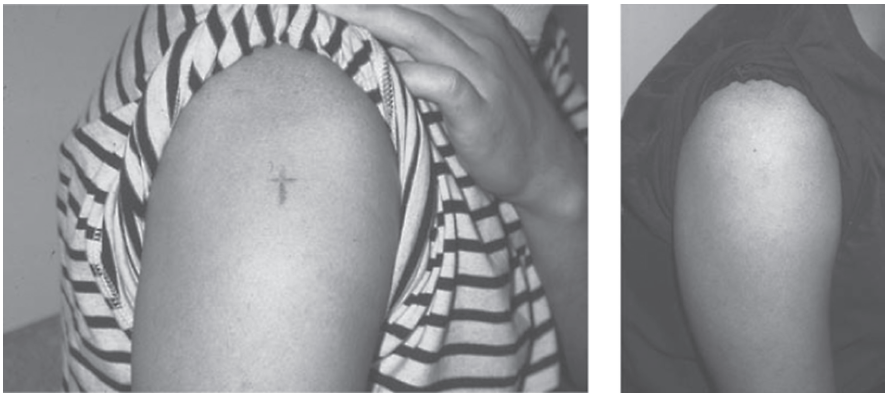

(a) (b)

FIGURE 19.4 (a) Amateur tattoo. (b) Appearance after one treatment with the Q-switched Nd:YAG laser. Copyright Springer.

is likely to be in the mid-follicle or isthmus region [15]. The exact follicular anatomy with respect to follicular stem cells and targeted chromophores remains unclear.

The epidermis also contains melanin and selective photothermolysis of hair follicles can occur either through differential amounts of the target chromophore, i.e. high melanin in dark-coloured hairs and low melanin in pale skin, or through appropriate selection of pulse duration. The thermal relaxation time of the epidermis can be considered to be around 3–10 ms. Therefore, pulsewidths matched to the hair follicle are more effective in causing selective thermal injury selectively to these structures. For the first reason given above, the best results with many lasers for selective thermolysis of hair have been in patients with white or pale skin and black hair. Further protection of the epidermis can be obtained by a variety of intra-operative cooling devices.

Successful hair removal in darker-skinned patients has generally been with diode and Nd:YAG lasers. The longer wavelengths emitted by these lasers are deeply penetrating and less well absorbed by epidermal melanin. Higher fluences are required to produce sufficient photothermal injury of pigmented hair and this must be accompanied by epidermal cooling.

Later an extended theory of selective photothermolysis has been proposed [16] which is of relevance to laser hair removal. This applies where the target is damaged by heat diffusion from a chromophore it does not contain. The thermal damage time is significantly longer than the thermal relaxation time as the temperature at the target must be sufficient to induce thermal injury. The power density should be high enough to raise the temperature but low enough to prevent chromophore destruction and subsequent absorption loss.

19.5 Carbon Dioxide and Erbium:YAG Lasers in Dermatology

The carbon dioxide (CO_2) laser is currently enjoying a renaissance with the advent of laser skin resurfacing. The CO_2 laser emits far-infrared light at 10 600 nm which is strongly absorbed by water, the major component of most human soft tissues. Thus, when the CO_2 laser impacts on skin it will be absorbed by this tissue superficially.

Early CO_2 lasers operating in the continuous mode could be shuttered to deliver 0.1–1 s pulses. The laser was used either in *focused* mode as an *excisional* laser or in *defocused* mode as an *ablative* laser. The ability of the CO_2 laser to excise tissue is a function of its irradiance. As can be seen from the following formula, irradiance is a function of the inverse of the *square* of the spot diameter.

$$\text{Irradiance}\,(\text{W cm}^{-2}) = \frac{\text{Power output (W)}}{\text{Laser beam cross-sectional area (cm}^{-2})}. \tag{19.2}$$

When the CO_2 laser is used for incisional surgery, its beam is focused to very small diameters with a relatively high output of 20–35 W. This results in extremely high irradiances of between 50 000 and 100 000 W cm^{-2}, which will incise tissue.

The response of tissue to CO_2 laser irradiation will be heating up to 100°C when vaporization occurs. When heating is rapid, all targeted tissue water is vaporized as steam. Tissue structures explode outwards owing to the rapid thermal expansion of water, shock waves and cavitation effects. This tissue combined with steam constitutes the plume of smoke arising from CO_2 laser-irradiated tissue and there is minimal thermal damage to adjacent tissue. This is due to heat diffusion from the impact site causing a zone of coagulative necrosis 0.6–1.3 mm away. If the fluence used is insufficient to vaporize tissue rapidly, coagulation, desiccation and carbonization will occur.

For ablative CO_2 laser use, the surgeon is attempting to remove controlled superficial layers of tissue with minimal thermal damage to surrounding tissue. The beam is defocused with a spot diameter of 0.5–5 mm; this larger beam will significantly reduce the irradiance. The ablation threshold for skin is greater than 5 J cm^{-2} so a defocused CO_2 laser must have a sufficient power output to deliver sufficient fluences within a short period of time.

With both excisional and ablative techniques, the width of the zone of thermal damage of adjacent tissue is directly proportional to the duration of the exposure. Even with good techniques, when a continuous-wave (cw) CO_2 laser is used to vaporize tissue, the residual thermal damage can be between 300 and 600 μm in depth. Short-pulsed CO_2 lasers have significantly reduced the depth of residual thermal damage.

To ablate tissue, fluences in excess of 5 J cm^{-2} are required. To minimize residual thermal damage, this energy should be delivered within the thermal relaxation time of the 30 μm layer of tissue absorbing the laser, which has been calculated to be less than 1 ms [17].

Super-pulsed lasers were developed with peak powers 2–10 times the power of cw CO_2 lasers. The laser produces pulse durations typically of 0.1–0.9 ms, with repetitions of 100–5000 Hz. By using a super-pulsed CO_2 laser high peak powers with short cooling pauses between each pulse are delivered. Using a pulsed CO_2 laser, it is possible to ablate tissue while reducing peripheral thermal damage to a thickness less than 100 μm.

The development of CO_2 lasers that could deliver sufficient energy to vaporize tissue within the thermal relaxation time (1 ms) *in a single pulse* substantially advanced the use of lasers for resurfacing photodamaged skin, including mild and moderate wrinkles.

19.5.1 Erbium:YAG Laser in Dermatology

The erbium:YAG (Er:YAG) laser is a laser with a wavelength of 2940 nm, which is in the mid-infrared invisible light spectrum. This wavelength corresponds to the main peak of water absorption (Figure 19.5). This wavelength has a 10–15 times greater water absorption than a CO_2 laser. Thus, the penetration depth into tissue of the Er:YAG laser beam is limited to a small volume of tissue. The ablation threshold of the Er:YAG laser for human skin has been calculated at 1.6 J cm^{-2} [18] compared with 5 J cm^{-2} for short-pulsed CO_2 lasers. Thus, with typical operating fluences of 5–15 J cm^{-2}, the Er:YAG laser is operated well above the ablation threshold which results in greater ablation and less residual thermal damage. Because of the short penetration depth of the 2940 nm radiation, the laser-heated

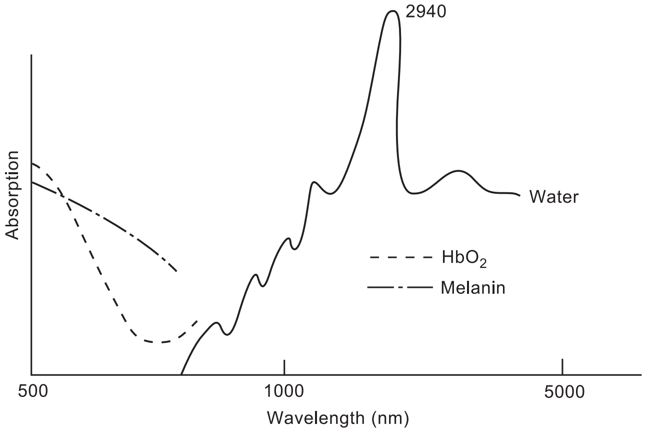


FIGURE 19.5

layer of tissue is initially 1 μm wide with a thermal relaxation time of approximately 1 μs. Thus single pulses of the Er:YAG less than 1 *μ*m in duration will minimize the zone of thermal damage.

Thus, the Er:YAG laser can ablate very narrow layers of tissue with minimal residual thermal damage. The potential advantages of this laser are precise control of tissue ablation with reduced thermal damage. Possible disadvantages are multiple passes to produce desired tissue ablation, no collagen contraction in resurfacing photodamaged skin and no haemostasis due to the very narrow zone of residual thermal damage.

REFERENCES

1. Anderson R R and Parrish J A 1983 Selective photothermolysis: precise microsurgery by selective absorption of pulsed radiation *Science* **20** 524–7.
2. Anderson R R and Parrish J A 1981 Microvasculature can be selectively damaged using dye lasers: a basic theory and experimental evidence in human skin *Lasers Surg. Med.* **1** 263–6.
3. Barsky S H, Rosen S, Geer D E and Noe J M 1976 The nature and evolution of port wine stains: a computer-assisted study *J. Invest. Dermatol.* **74** 154–7.
4. van Gemert M J C, Welch A J, Pickering J W, Tan O T and Gijsbers G H M 1995 Wavelengths for laser treatment of port wine stains and telangiectasia. *Lasers Surg. Med.* **16** 147–55.
5. Dierickx C C, Casparian J M, Venugopalan V, Farinelli W A and Anderson R R 1995 Thermal relaxation of port wine stain vessels probed *in vivo*: the need for 1–10 millisecond laser pulse treatment *J. Invest. Dermatol.* **105** 709–14.
6. Garden J M et al 1986 Effect of dye laser pulse duration on selective cutaneous vascular injury *J. Invest. Dermatol.* **87** 653–7.
7. Tan O T, Kerschmann R and Parrish J A 1984 The effect of epidermal pigmentation on selective vascular effects of pulsed laser *Lasers Surg. Med.* **4** 365–74.
8. Nelson J S et al 1995 Dynamic epidermal cooling during pulse laser treatment of port wine stain *Arch. Dermatol.* **131** 695–700.
9. Keijzer M, Pickering J W and van Gemert M J C 1991 Laser beam diameter for port wine stain treatment *Lasers Surg. Med.* **11** 601–5.
10. Lucassen G W, Verkruyesse W, Keijzer M and van Gemert M J C 1996 Light distributions in a port wine stain model containing multiple cylindrical and curved blood vessels *Lasers Surg. Med.* **18** 345–57.
11. Tan O T, Murray S and Kurban A K 1989 Action spectrum of vascular specific injury using pulsed irradiation *J. Invest. Dermatol.* **92** 868–71.
12. Ross V, Naseef G, Lin G, Kelly M, Michaud N, Flotte TJ, Raythen J and Anderson R R 1998 Comparison of responses of tattoos to picosecond and nanosecond Q-switched neodymium: YAG lasers *Arch. Dermatol.* **134** 167–71.
13. Baumler W, Eibler E T, Hohenleutner U, Sens B, Sauer J and Landthaler M 2000 Q-switch laser and tattoo pigments: first results of the chemical and photophysical analysis of 41 compounds *Lasers Surg. Med.* **26** 13–21.
14. Lin T-Y D et al 1998 Hair growth cycle affects hair follicle destruction by ruby laser pulses *J Invest. Dermatol.* **111** 107–13.
15. Kim J-C and Choi Y-C 1995 Hair follicle regeneration after horizontal re-sectioning: implications for hair transplantations *Hair Replacement: Surgical and Medical* ed D B Stough and R S Haber (St Louis: Mosby-Yearbook) pp 358–63.
16. Altshuler G B, Anderson R R, Manstein D, Zenzie H H and Smirnov M Z 2001 Extended theory of selective photothermolysis *Lasers Surg. Med.* **29** 416–32.
17. Walsh J J, Flotte T J, Anderson R R and Deutsh T F 1998 Pulsed CO$_2$ laser tissue ablation: effect of tissue type and pulse duration on thermal damage *Lasers Surg. Med.* **8** 108–18.
18. Hohenleutner U, Hohenleutner S, Baumler W and Landthaler M 1997 Fast and effective tissue ablation rates and thermal damage zones *Lasers Surg. Med.* **20** 242–7.

FURTHER READING

Ross E V, McKinlay J R and Anderson R R 1999 Why does carbon dioxide resurfacing work? *Arch. Dermatol.* **135** 444–54.

In depth review of mechanisms of laser induced photo-rejuvenation.

2000 *Lasers in Surgery and Medicine* **26**(2).

Special issue focused on areas of dermatological laser treatments with greatest clinical potential.

Lanigan S W 2000 *Lasers in Dermatology* (London: Springer).

Compact up-to-date textbook on all aspects of cutaneous laser therapy.

Steiner R, Kaufmann, Landthaler M and Braun-Flaco O (ed) 1989 *Lasers in Dermatology* (Berlin: Springer).

Symposium proceedings.

20

Therapeutic Applications: Lasers in Vascular Surgery

Mahesh Pai

CONTENTS

20.1 Atheroma and Angioplasty-Like Injury 273
20.2 Laser Angioplasty 274
20.3 Reported and Ongoing Studies on Excimer Laser Angioplasty 275
  20.3.1 Coronary Arteries 275
  20.3.2 Peripheral (Femoral) Arteries 275
20.4 PDT in the Arteries 276
  20.4.1 PDT of Injured and Diseased Arteries 276
  20.4.2 Light Delivery 276
  20.4.3 Clinical Data 277
20.5 Conclusions 278
References 278

20.1 Atheroma and Angioplasty-Like Injury

Atheromatous occlusive disease is the major cause of mortality and morbidity in the western world. It was responsible for just under 1 million deaths (39% of all deaths) in the United States in 2000 and just over 210 000 deaths in the UK in 2001 (40% of all deaths). The disease is diffusely distributed throughout the arterial tree. The arterial lumen is narrowed or occluded by an intimal plaque made up of foamy macrophages (a type of inflammatory cell filled with fat), free cholesterol and fibrous tissue. Focal accumulations of these plaques lead to a discrete narrowing or an occlusion in arteries such as the coronary and the carotid and give rise to manifestations such as angina, while acute ruptures of these plaques along with thrombosis (clotting within the artery) lead to myocardial infarcts (death of muscle tissue in the heart) and strokes. The narrowed arteries are amenable to surgical bypass grafting or procedures such as percutaneous transluminal angioplasty (PTA, a technique for opening up the block in an artery by sliding a device along the arterial tree to the site of the block from a convenient access point).

PTA is a relatively simple procedure (Figure 20.1). A needle is inserted into a point where a large artery is close to the skin, usually the femoral artery in the groin. Under X-ray guidance, a fine wire is introduced through the needle into the arterial lumen and guided to and passed through the narrowed segment of the artery. A long canula (thin, flexible plastic tube) is passed over the guide wire and an angiogram is taken to confirm the narrowing. This involves the injection of a radio-opaque dye through the canula and taking X-rays of the involved segment of artery. The canula is then removed and replaced by an undilated balloon catheter, which is slid over this guide wire. Dilatation of the balloon opens up the narrow segment and pushes the atheroma into the wall, reopening the lumen to blood flow. If the lesion is a narrowing one, then the procedural success rate is about 90%–95%. However, this rate falls dramatically in the case of complete occlusions. This may be due to the inability to pass a guide wire through the occlusion or because a balloon or a sheath cannot be threaded over the guide wire.

Even when the initial PTA is successful, unfortunately, 30%–50% of these patients have recurrent arterial narrowing in the first few months, known as restenosis. Most failures are due to exuberant healing of the balloon catheter-induced injury, especially by proliferation of vascular smooth muscle cells (VSMC) in the arterial wall, a process known as neointimal hyperplasia (NIH). Other mechanisms include deposition of extracellular matrix and shrinkage in the dimensions of the artery (negative arterial remodelling).

Almost a million coronary angioplasties are carried out in the USA every year and about half of them involve the placement of a stent, a slotted metal mesh tube, inside the artery after the PTA (Figure 20.2). In Europe that proportion reaches 80%. The support the stent affords the vessel wall reduces immediate recoil and helps stabilize the torn wall, so the immediate results are better than simple PTA. The excellent gain in lumen size means that the number of patients developing clinically significant restenosis is reduced after stenting in comparison to PTA alone. Only 10%–20% of patients require re-intervention after stenting. Yet stenting does not prevent NIH. An aggressive neointimal response is stimulated but with the greater lumen size afforded by the stent, more NIH can be accommodated; hence, the relatively lower rate of clinical restenosis when stenting is compared with PTA. However, when restenosis does occur, its treatment is very difficult.

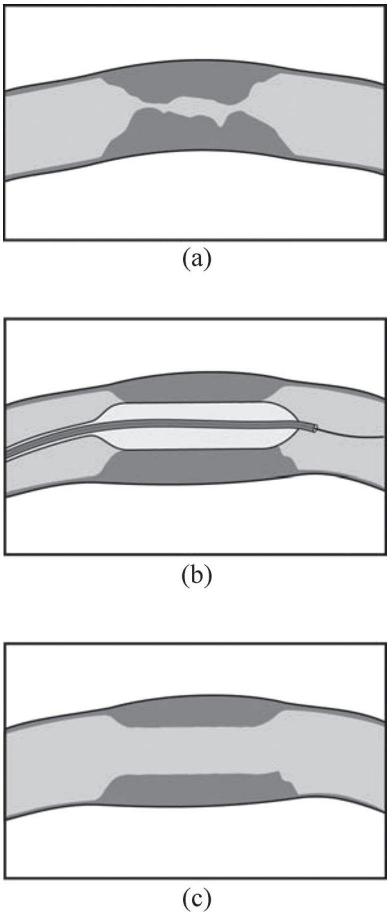


FIGURE 20.1 Angioplasty. (a) A tight stenosis in an artery narrowing the lumen by >75%. (b) Under X-ray guidance a guide wire is negotiated through the stenosis and a balloon is passed over it and aligned with the lesion. The balloon is inflated and the narrowing widened. (c) The balloon is deflated and withdrawn along with the guide wire leaving a lumen with no significant stenosis.

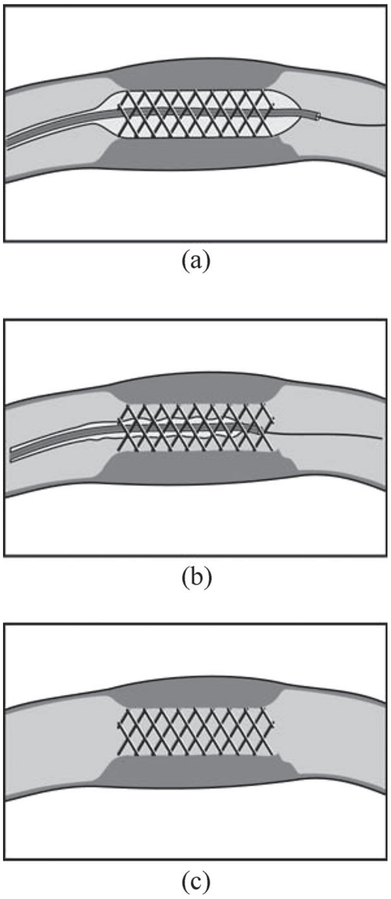


FIGURE 20.2 Stenting. (a) An unexpanded stent is pre-mounted on an angioplasty balloon and introduced on a guide wire and aligned with the lesion. On inflation of the balloon, the stent expands and is embedded in the lesion after angioplasty. (b) The balloon is deflated and withdrawn along with the guide wire. (c) The stent increases the diameter of the vessel. Hence, a thicker neointima is needed to cause significant restenosis.

Lasers have been used in the treatment of atherosclerotic disease in the last two decades. When lasers were first available for use in medicine there was considerable excitement about their potential use in arterial disease for relieving stenosis. This initial enthusiasm has since been somewhat tempered. However, in the last decade, the use of lasers in photodynamic therapy (PDT) as an adjunct to other techniques for relieving arterial stenosis has re-stimulated interest. PDT could develop an important role in the prevention of restenosis.

20.2 Laser Angioplasty

Attempts have been made to use lasers as an aid to PTA since the 1980s. In 1963, only three years after the development of lasers, laser energy was used to vaporize an atherosclerotic plaque. However, it was not until 1983 that Choy performed the first endovascular (i.e. within the vessel lumen) angioplasty using an argon laser, initially in animal models but later in a pilot clinical study. Lack of proper equipment and a poor understanding of dosimetry led to a number of complications including perforation of the artery, acute occlusion and critical ischemia (a condition in which blood supply to the organ is significantly reduced causing irreversible damage such as gangrene of the foot or loss of cardiac muscle). The system was tried in both the coronary and peripheral circulation with an equal lack of success, although in the late 1980s and early 1990s, better understanding led to some improvement in safety and efficacy.

Introduction of laser energy into the lumen of an artery can have three different effects: thermal damage, photoablation, i.e. tissue removal by bond breakage of the peptide bonds in organic molecules, or photodisruption, i.e. tissue disruption by high-peak-power laser pulses [1]. It is preferable that thermal effects are minimized. Different lasers such as excimer lasers and holium lasers have been used as they can vaporize plaques without thermal damage to the underlying tissue. Using pulsed lasers with cooling with saline infusions further reduced the risk of thermal damage. Use of optical fibres allows the manipulation of a laser catheter over a guide wire under X-ray guidance for alignment with the lesion to be treated. The laser catheter is advanced slowly (0.2–0.5 mm s^{-1}) and stopped when the plaque ends. Multi-fibre laser catheters were

TABLE 20.1

Lesion Types	Contraindications
Debulking of plaque before stenting	Vessels too small for the catheter
Total occlusions	Tortuous vessels
In-stent restenosis	Lesion which cannot be transversed by a guide-wire[a]
Thrombosis	
Restenosis in vein grafts	Dissection or perforation of the vessel pre-laser
Lesion which cannot be passed by a balloon	

[a] This used to be an indication but is no longer considered safe.

developed to enhance flexibility of the catheter, so decreasing the likelihood of a perforation. This has also allowed larger diameter laser catheters (up to 2.5 mm in diameter) to be used. These changes certainly improved the results of laser angioplasty, and in 1993, the American FDA (Food and Drug Administration) approved the use of pulsed excimer laser systems for clinical use in coronary vessels. Ultraviolet excimer lasers provide light at 308 nm. Lasing is initiated in 5 s trains (125 pulses at the frequency of 25 Hz), while saline is flushed through the guide-wire channel. At this rate the penetration depth is 35–50 µm. The energy fluence is typically 40 mJ mm^{-2}, but this can be increased up to 60 mJ mm^{-2}. Since then a number of different trials have studied the use of this laser. The indications for which laser angioplasty has been used are summarized in Table 20.1 [2].

20.3 Reported and Ongoing Studies on Excimer Laser Angioplasty

20.3.1 Coronary Arteries

Köster et al. [3] reported the use of laser angioplasty with adjuvant balloon PTA in 1999. Four hundred and forty patients received excimer laser ablation with balloon PTA for in-stent restenosis with a procedural success rate of 91%. They used laser catheters varying in size from 1.4 to 2.0 mm diameter. The maximum fluence they used was 49 ± 5 mJ mm^{-2} at 26 ± 4 Hz. The laser was used to cut a hole through the lesion, and the balloon was then used to dilate the treated segment a bit more to end up with a lumen at least 3 mm in diameter. Procedural complications included dissection (splitting of the layers in the wall of the artery) in 4.8%, perforation in 0.9% and acute re-occlusion in 2%. They recorded a myocardial infarction rate of 3.2%, which is comparable to standard PTA for in-stent restenosis. A multi-centre, randomized controlled trial is currently under way in the USA.

The ERBAC study, i.e. excimer laser (E) *versus* rotational atherectomy (R) *versus* balloon angioplasty (BA) trial, was reported by Reifart et al. [4]. Atherectomy is the process of cutting through the plaque using a mechanical device such as a miniature rotating blade. Six hundred and eighty-five patients with symptomatic complex coronary lesions were randomly allotted to each group with the primary end point being the procedural success. They used laser catheters varying in size from 1.3 to 2.2 mm. The fluence varied from 45 to 70 mJ mm^{-2} at 20–30 Hz. While the procedural success rate was highest in the rotational atherectomy group (89%) followed by the PTA group at 80% and the laser group (77%), at 6 months the PTA group had the lowest restenosis rate at 47% compared to 59% in the laser group and 57% in the rotational atherectomy group.

The LAVA study compared the use of laser angioplasty using a holmium YAG laser as an adjuvant to balloon PTA with plain-balloon PTA with no difference seen in procedural success, complication or restenosis rates at 11 months of follow-up. Similar results were also seen in the AMRO trial designed in the same way as the LAVA trial except that an excimer laser was used [5].

Non-randomized trials have shown the benefit of laser angioplasty in blocked saphenous vein grafts (which are used as a conduit when doing surgical arterial bypasses such as a coronary artery bypass) but as yet no randomized study has confirmed this benefit.

There is also the possibility of using lasers to vaporize thrombi in the emergency situation just after a coronary artery gets occluded. However, this technique still needs refinement before it can be put into clinical practice. Currently, when a patient is admitted to hospital with a heart attack, it is a standard practice to consider the patient for thrombolysis, i.e. dissolution of clot within the arterial lumen using drugs. This is effective in reducing both the morbidity (complications) and mortality associated with the infarct. Use of the excimer laser is currently being assessed as an alternative to these drugs as thrombolytic drugs carry the risk of causing bleeding elsewhere in the body, particularly the brain.

20.3.2 Peripheral (Femoral) Arteries

Two ongoing trials are looking at the use of excimer laser angioplasty (ELA) in the femoral arteries in the top of the legs. The PELA or peripheral ELA trial is a randomized, multicentre trial comparing safety and effectiveness between simple PTA and laser angioplasty in long (>10 cm) lesions in the superficial femoral artery (SFA). The SFA is the main artery carrying blood to the leg and foot and is usually blocked in the thigh. Laser catheters ranging in size from 1.4 to 2.5 mm in diameter are being used for light delivery. The primary end point is patency measured by Duplex (ultrasound scanning) at 12 months with clinical follow-up at 1, 3, 6 and 12 months [6].

The second ongoing trial is the LACI trial looking at the use of laser angioplasty in critical ischemia. This is a multicentre feasibility study to assess the safety and effectiveness of laser angioplasty in vessels below the knee in limb-threatening ischemia. As these vessels are extremely small, PTA does not work here. If some revascularization is not attempted or possible then the limb will be amputated. The primary end point is limb-survival at 3 months and patients are being followed up for 6 months. The phase 1 study with 25 patients has been completed and a phase 2 study has commenced [6].

Despite the initial failure of laser angioplasty to fulfil its early promise in atherosclerotic disease, its use has been refined later. With appropriate case selection, utilization of equipment, understanding of its biology and application of safe lasing techniques, it is beginning to receive growing interest. It could prove a vital tool in this wide-spread clinical problem.

20.4 PDT in the Arteries

PDT involves accumulation of a photosensitizing drug by cells or tissues, which is then activated by illumination with non-thermal laser light in the presence of oxygen. This results in the generation of reactive oxygen species, which cause localized cell death. It is being developed for the treatment of a variety of malignancies. Experiments designed to understand the effects of PDT on blood vessels lead to the realization that PDT might be of value in the treatment of vascular disease.

Several different photosensitizers have been shown to accumulate in the walls of blood vessels. These include haematoporphyrin derivative (HpD), 5-aminolaevulinic acid (ALA), chloroaluminium sulphonated phthalocyanine (CASP), motexafin lutetium (Antrin®) and mono-l-aspartyl-chlorin e6 (NPe6). The drugs can be identified within all layers of the normal arterial wall and also within the macrophage-rich plaques of atherosclerosis induced experimentally in rabbits by cholesterol feeding. The rate of accumulation of drug in the wall of the blood vessel depends on the characteristics of the particular photosensitizer and the route of delivery, which may be by mouth, intravenously (IV) or by local delivery directly into the artery. For example, NPe6 concentration in the arterial wall peaks within 1 h of IV administration [7], while 5-ALA is a pro-drug, which must be converted to its active metabolite, protoporphyrin-IX, a photosensitizing chemical, which is a precursor in the natural synthesis of haem. It takes 4–6 h for the active photosensitizer to accumulate in the arterial wall after oral administration [8]. The arterial wall concentration of motexafin lutetium rises more slowly, peaking in the media layer after 24 h [9]. It is also taken up by the cholesterol-rich foamy macrophages (a type of white blood cell). Different kinetics of accumulation in different layers of the blood vessel wall, when combined with timely light delivery to activate drug, may allow some targeting of specific layers of the artery wall.

Laboratory studies in small rodents have shown that in very small blood vessels, there is rapid arrest of the circulation due to damage to the lining of the vessel. However, in all larger vessels which have connective tissue (collagen) in the wall of the artery, the blood flow is maintained at all times, even though the cells lining the artery and the muscle cells in the arterial wall are killed. This pattern is reproduced after activation of all the sensitizers tested in the vascular system [7, 9–11]. Applications of PDT to the normal arteries of larger animals such as the pig, whose coronaries and peripheral resemble those of humans, have duplicated the results seen in small animals. At 3 days after PDT, there is almost complete depletion of endothelial (lining) and medial cells in porcine coronary and iliac arteries. As mentioned earlier, these are the very cells involved in the process of restenosis after PTA [8]. Hence, the potential for using PDT as an adjuvant therapy with PTA to prevent restenosis.

20.4.1 PDT of Injured and Diseased Arteries

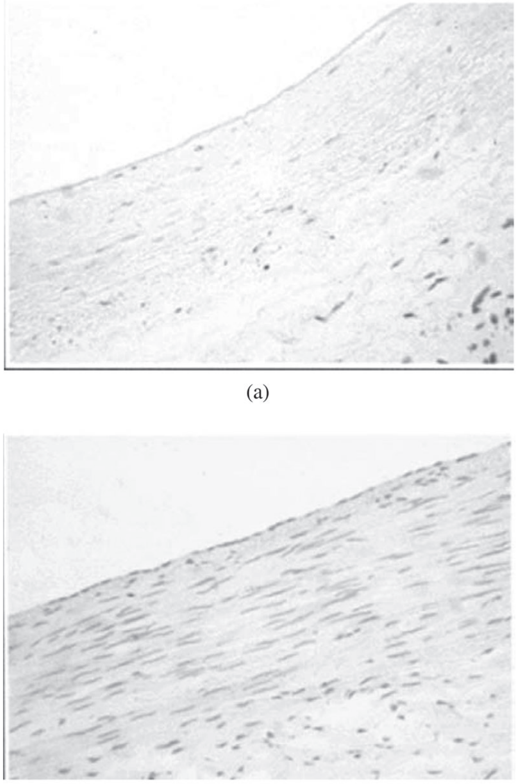
In vitro (cell culture) inhibition of VSMC proliferation using HpD has been recognized for many years. Several groups have reported that PDT has an inhibitory effect on NIH in small animal models [7, 10–12]. Angioplasty-like balloon injury of an artery invariably stimulates NIH. In the rat model of balloon dilatation, longitudinal traction and torsion combine with circumferential stretch to injure the artery. PDT with any of a number of drug–light combinations has been shown to abolish NIH effectively in the balloon-injured rat carotid artery. The media of the treated area are rendered acellular (i.e. all the cells are killed), and although the endothelium re-grows, the neointima is strikingly absent with a sharp demarcation line at the edge of the treated area. As well as preventing intimal hyperplasia (NIH) after balloon injury of the rat carotid, PDT can also treat restenosis in this model, causing the intimal hyperplasia to disappear rapidly.

The larger animal models of angioplasty such as the pig are more akin to the clinical scenario in that the injury is focal and involves stretch and rupture of the elastic laminae of the artery wall. The endothelium is destroyed by the angioplasty procedure itself but the adjuvant PDT in this situation leads to rapid loss of VSMC so that the media are acellular within 3 days of the treatment (Figure 20.3). Just as in the small animal models, there is a reduction in the neointimal response to angioplasty but a contribution to the lumen enlargement associated with PDT comes from preservation of the dimensions of the injured artery segment, i.e. inhibition also of the negative vascular remodelling [13] (Figure 20.4).

PDT has been shown to kill the VSMC and endothelium by apoptosis. This is a process by which cells appear to kill themselves without exciting an inflammatory response in the surrounding tissue. This is responsible for the lack of scarring in PDT-treated arteries. An important observation was that the acellular carotid artery of the rabbit maintains its structural integrity and had a higher than normal bursting pressure 3 days after PDT. In experimental atheromatous plaque, PDT disrupts the architecture of the elastin network. This may reduce the elastic recoil following angioplasty [14]. It is clear that PDT has the exact effects that are needed in the arterial wall to facilitate its use as an adjuvant to prevent restenosis after PDT.

20.4.2 Light Delivery

Endovascular light delivery is the delivery of light within the lumen of the artery under X-ray guidance. Endovascular light delivery catheters can be devised by modification of existing angioplasty equipment (Figure 20.4). Light can be delivered via optical fibres with 100–200 μm core diameter, a size similar to that of the guide wire within an angioplasty balloon. Newer diode laser systems are much more portable than the



(a)

(b)

FIGURE 20.3 Photomicrographs of pig coronary arteries, 3 days after balloon injury: (a) balloon injury alone, showing marked cellularity in the intima and media of the artery; (b) and adjunctive PDT (60 mg kg^{-1} ALA given 5–7 h prior to delivery of 50 J cm^{-2} light at 635 nm). The light was delivered immediately after the balloon injury. There is almost complete loss of all cells from the intima and media. (Reproduced with permission from Ref. [8].) Haematoxylin and eosin stain.

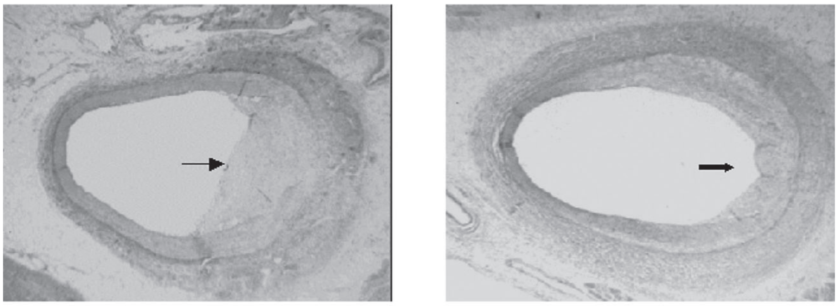
older pumped dye models. The fluence rate of the light on the arterial wall must be kept below 120 mW cm^{-2} to avoid thermal effects. Alternatives such as miniature light-emitting diode systems could allow light generation inside the artery and be both flexible and robust. If the wavelength for drug activation is towards the visible red, then either of these endovascular light-delivery approaches will demand that blood is transiently excluded from the lumen of the artery during illumination. Modification of existing balloon catheters to facilitate light delivery may achieve this aim.

A further consideration for endovascular approaches is that of depth of tissue effects. The tissue penetration is largely determined by the wavelength of the light used, which in turn is dictated by the sensitizing drug employed. The vascular target in the heart is the large epicardial arteries, which lie on the surface of the heart. The ideal drug for cardiac application requires an activating light that penetrates through the wall of the artery but only just reaches the actual surface of the heart. This is a depth of less than 2 mm even for the large coronary arteries close to their origin from the aorta, just where it leaves the heart. However, there is the danger of incomplete coverage of the treatment area. There is evidence that balloon-injured areas which are untreated or undertreated with PDT may develop more NIH, so the complete field of injury related to the angioplasty needs to be covered.

20.4.3 Clinical Data

Studies of PDT on vascular disease in patients remain very limited. Only two clinical studies have been published. Both are pilot studies to assess the safety and feasibility of the technique. Jenkins et al. [15] described seven patients (eight lesions), all of whom had had previous angioplasty for symptomatic SFA disease but had developed further symptoms. Five of the eight lesions had severe restenosis while three had complete occlusions. All patients underwent repeat angioplasty of the SFA but this time with adjuvant PDT. 5-ALA was used as the photosensitizing agent and the artery was illuminated endovascularly (with blood excluded by the balloon catheter) with 50 J cm^{-2} of 635 nm light from a copper vapour-pumped dye laser. At the 6 months follow-up, all of these patients were free of symptoms. They all had angiograms at 6 months which confirmed the absence of restenosis. These patients have been followed up for up to 4 years and the results remained encouraging. Only one lesion has restenosed at the treated site (Figure 20.5) [15].

Rockson et al. [16] reported a phase I study in the use of PDT with motexafin lutetium® in lower limb arteries including the iliac, common and superficial femoral. In a two-part experimental design, symptomatic patients were divided into two groups. The first group was given an ascending dose of



(a) (b)

FIGURE 20.4 Photomicrographs of pig coronary arteries, 28 days after balloon injury: (a) balloon injury alone, showing marked fibrocellular NIH (arrow); (b) and adjunctive PDT (100 mg kg^{-1} ALA given 5–7 h prior to delivery of 50 J cm^{-2} light at 635 nm). The light was applied immediately after the balloon injury. There is significant inhibition of NIH. (Reproduced with permission from Ref. [13].) Haematoxylin and eosin stain.

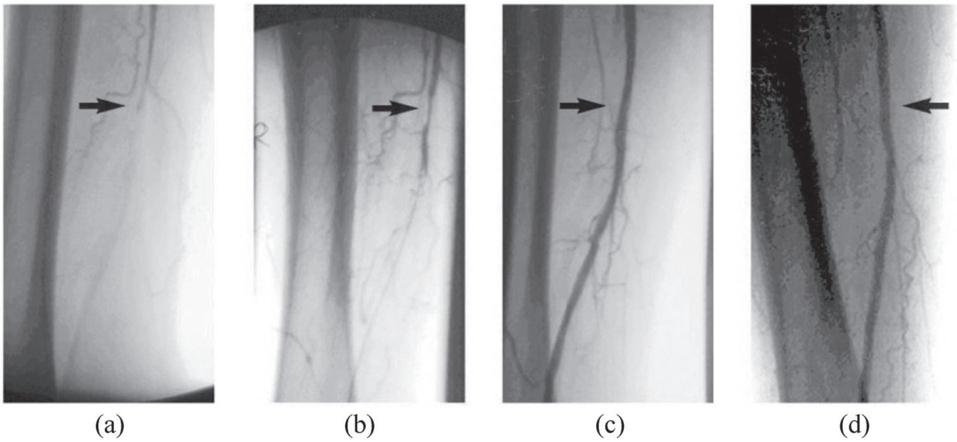


FIGURE 20.5 Right femoral angiograms illustrating (a) occlusion of the femoral artery before angioplasty and PDT, (b) guide wire successfully across the occlusion, (c) angiographic appearance after angioplasty and adjunctive PDT and (d) repeat angiogram of the right femoral artery 26 months later demonstrating patency and no evidence of restenosis.

motexafin lutetium with a standard light dose of 400 J cm^{-1}. All doses ranging from 1–5mgkg^{-1} were well tolerated with only minor side-effects of self-limiting skin photosensitivity. The second group of patients were given 2–4 mg kg^{-1} and light doses of 500, 625 and 781 J cm^{-1} for each drug dose. Motexafrin lutetium was given intravenously and the vessels illuminated 24 h later. No PTA was done and the wavelength selected for drug activation avoided the need for blood exclusion, so light diffusing from a bare laser fibre was used for illumination. Again all light and drug doses were well tolerated with only minor cutaneous photosensitivity being seen. Although not designed to test clinical efficacy, there was a small (median 4%) decrease in arterial stenosis and a small improvement in median ankle brachial pressure index (a measure of peripheral blood flow) of 3.6%. Early results of motexafin lutetium PDT as an adjuvant to coronary stenting suggest that it is well tolerated [16].

PDT clearly affects the regulation of vascular cell proliferation after injury significantly. It causes a temporary depletion of endothelial and medial cells in the arterial wall when used as an adjuvant to angioplasty. In an environment conducive to proliferation and neointima formation, it effectively prevents this. However, unlike brachytherapy (endovascular radiation—another technique which is under investigation for treatment of restenosis), this depletion is not permanent. Rapid endothelial and delayed medial recovery prevents any thrombotic complications and preserves structural integrity in the long term.

20.5 Conclusions

Angioplasty and stenting procedures are of increasing popularity in the treatment of coronary and peripheral vascular disease but restenosis remains a problem in a significant proportion of patients. The increasing refinement of light delivery systems for PDT as an adjuvant to PTA could mean that it carves a niche for itself in the treatment of occlusive arterial disease. PDT is a promising option as the experimental studies and preliminary clinical data available suggest that it is feasible, safe, relatively straightforward and likely to be effective as well as having less long-term effects than brachytherapy. PDT may also be of value for treating restenosis. There is now enough evidence for the potential of PDT for these indications to justify larger scale clinical trials.

REFERENCES

1. Deckelbaum LI 1994 Coronary laser angioplasty *Lasers Surg. Med.* **14** 101–10.
2. Topaz O et al. 2001 Excimer laser revascularisation: current indications, applications and techniques *Lasers Med. Sci.* **16** 72–7.
3. Koster R et al. 1999 Laser angioplasty of restenosed coronary stents: results of a multicenter surveillance trial. The laser angioplasty of restenosed stents (LARS) investigators *J. Am. Coll. Cardiol.* **34** 25–32.
4. Reifart N et al. 1997 Randomized comparison of angioplasty of complex coronary lesions at a single center. Excimer laser, rotational atherectomy, and balloon angioplasty comparison (ERBAC) study *Circulation* **96** 91–8.
5. Stone GW et al. 1997 Prospective, randomized, multicenter comparison of laser-facilitated balloon angioplasty versus stand-alone balloon angioplasty in patients with obstructive coronary artery disease. The laser angioplasty versus angioplasty (LAVA) trial investigators *J. Am. Coll. Cardiol.* **30** 1714–21.
6. Tcheng J E and Volkert N 2001 Current multicentre studies with the excimer laser: design and aims *Lasers Med. Sci.* **16** 122–9.
7. Nagae T et al. 2001 Endovascular photodynamic therapy using mono-L-aspartyl-chlorin e6 to inhibit Intimal hyperplasia in balloon-injured rabbit arteries *Lasers Surg. Med.* **28** 381–8.
8. Jenkins M P et al. 1998 Intra-arterial photodynamic therapy using 5-ALA in a swine model *Eur. J. Vasc. Endovasc. Surg.* **16** 284–91.

9. Rockson S G, Lorenz D P, Cheong W F and Woodburn K W 2000 Photoangioplasty: an emerging clinical cardiovascular role for photodynamic therapy *Circulation* **102** 591–6.
10. Nyamekye I et al. 1996 Inhibition of intimal hyperplasia in balloon injured arteries with adjunctive phthalocyanine sensitized photodynamic therapy *Eur. J. Vasc. Endovasc. Surg.* **11** 19–28.
11. Nyamekye I et al. 1995 Photodynamic therapy of normal and balloon-injured rat carotid arteries using 5-amino-levulinic acid *Circulation* **91** 417–25.
12. Ortu P, LaMuraglia G M, Roberts W G, Flotte T J and Hasan T 1992 Photodynamic therapy of arteries. A novel approach for treatment of experimental intimal hyperplasia *Circulation* **85** 1189–96.
13. Jenkins M P et al. 2000 Reduction in the response to coronary and iliac artery injury with photodynamic therapy using 5-aminolaevulinic acid *Cardiovasc. Res.* **45** 478–85.
14. Grant W E et al. 1995 The effect of photodynamic therapy on the mechanical integrity of normal rabbit carotid arteries *Laryngoscope* **105** 867–71.
15. Jenkins M P et al. 1999 Clinical study of adjuvant photodynamic therapy to reduce restenosis following femoral angioplasty *Brit. J. Surg.* **86** 1258–63.
16. Rockson S G et al. 2000 Photoangioplasty for human peripheral atherosclerosis: results of a phase I trial of photodynamic therapy with motexafin lutetium (Antrin) *Circulation* 102 2322–4.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

21

Therapeutic Applications: Free-Electron Laser

**E. Duco Jansen, Michael Copeland, Glenn S. Edwards, William Gabella,
Karen Joos, Mark A. Mackanos, Jin H. Shen, and Stephen R. Uhlhorn**

CONTENTS

21.1 Introduction 281
21.2 Laser Ablation 282
  21.2.1 Ablation with the FEL 282
21.3 Towards Clinical Application of the FEL 283
  21.3.1 Beam Transport and Alignment 284
  21.3.2 Surgical Beam Delivery 284
21.4 Clinical Experience 285
  21.4.1 Neurosurgery 285
  21.4.2 Clinical Experience: Ophthalmology 285
21.5 Conclusions 287
References 287

21.1 Introduction

The holy grail of therapeutic laser applications in most clinical settings is to ablate the target tissue efficiently while minimizing collateral damage to adjacent tissue structures. Historically, this type of medical application of laser technology has largely been driven by empirical research in an effort to determine the mechanisms and results of aiming a given laser, often one that happened to be available, at a specific type of tissue and observing both (bio)physical interaction processes and histopathological results. More often than not, the lasers used were far from ideal in terms of their wavelength or pulse duration.

The free electron laser (FEL) has unique properties, most notably its wavelength tunability that allows access to spectral areas not covered by conventional lasers or where conventional lasers have insufficient power to ablate. This makes the FEL an attractive option to identify optimal laser parameters for specific and selected medical procedures and to study laser–material (tissue) interaction. Importantly, and in contrast to conventional lasers, the pulse duration/structure of the FEL is relatively independent of wavelength. One of the most significant outcomes of the basic biophysical and clinical studies with the FEL is that wavelengths in the mid-IR part of the spectrum, specifically at 6.45 μm where tissue amide-II bonds readily absorb the light, the FEL has superior cutting characteristics in soft biological tissue with <40 μm of collateral damage typically reported [1]. To date no other laser has been able to produce this wavelength with sufficient peak and average power to ablate tissue efficiently. However, the FEL is neither cheap nor easy to operate. Moreover, the mere thought of combining an FEL with a surgical suite and using the FEL as a clinical laser is daunting. While there are a number of FEL facilities in the world, Vanderbilt's W M Keck Free Electron Laser Center is the only FEL which integrates patient care, animal experimental facilities, physics and cell biology laboratories in the same building as the FEL itself [2,3]. Over the past decade, it has been shown that the FEL has the capability to ablate or cut both soft and hard tissue efficiently with minimal collateral damage [1,4–10]. Much research has been performed on the underlying biophysics of infrared ablation with the FEL [11–17]. Simultaneously translational research has been carried out that literally took the FEL from the optical bench to the operating room (OR) where experimental surgical procedures have been carried out successfully with the FEL and for which the food and drug administration (FDA) has granted an Investigational Device Exemption (IDE). However exciting this may be, these applications are limited to one or possibly a few specialized centres and are not easily translated to widespread clinical use. The working model of the FEL Center's medical research efforts has always been and must continue to be to use the FEL to identify the optimal parameters for a specific medical procedure, show proof-of-principle upon which dedicated, small and portable 'conventional' lasers (either gas lasers such as the strontium-vapour laser or solid state devices such as optical parametric oscillators (OPO's)) can be based that are amenable to widespread clinical use (see Figure 21.1). In this chapter the use of the FEL as a clinical laser will be discussed. In particular, the unique cutting characteristics of some of the mid-IR wavelengths (most notably 6.45 μm) will be discussed as well as the unique engineering challenges of clinical beam delivery to a patient some 100 m away from the laser at a wavelength and with peak irradiances where conventional fibre optics do not work. Finally, the initial clinical experiences will be discussed.

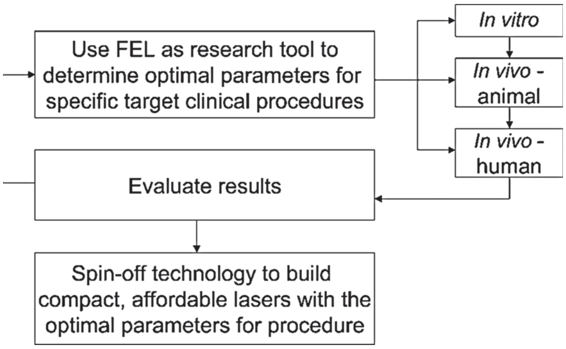

Use FEL as research tool to determine optimal parameters for specific target clinical procedures
In vitro
In vivo - animal
In vivo - human
Evaluate results
Spin-off technology to build compact, affordable lasers with the optimal parameters for procedure

FIGURE 21.1 Schematic diagram of the role of the FEL in therapeutic laser medicine.

21.2 Laser Ablation

Since the early days of lasers, researchers have explored their potential as medical devices. In this chapter we will ignore the tremendous impact lasers are having in diagnostic applications, sensing and imaging, and limit our discussion to therapeutic applications of laser technology, in particular with the FEL. Laser therapeutic use is photochemical, photothermal or photomechanical in nature. Ablation, the physical removal of tissue, is typically driven by photothermal interactions whereby the light is distributed in tissue governed by the tissue optical properties and subsequently absorbed upon which it is converted to thermal energy. Depending on the tissue optical properties, the depth of penetration of light may be deep (several cm) or very shallow (~1 μm). Generally speaking, for efficient tissue ablation, wavelengths that are strongly absorbed (i.e. UV or IR) are desirable [18–23]. Once the energy is deposited in the tissue, the subsequent thermal and thermomechanical processes are governed by the tissue thermal properties while the laser irradiance as well as rate of energy deposition have a major influence as well. From a clinical perspective, there are applications in which the combined cutting and coagulation with the laser is desirable. However, surgery in delicate areas of the body such as the eye and the brain requires precise cutting tools capable of efficient soft-tissue resection that minimize collateral damage to important neighbouring structures. Moreover, while conventional lasers (e.g. CO_2) sterilize wounds and provide haemostasis, thermal damage impairs optimal wound healing, may cause significant scar formation and can compromise the tensile strength of the wound [4,24–29]. The transition from cw lasers to pulsed lasers for ablative applications significantly reduced collateral damage by producing a thermal event that is shorter than the thermal relaxation time of the tissue and by delivering sufficient energy to ablate tissue with each pulse which removes hot material before heat is transferred to surrounding areas. The price paid for reducing thermal damage by shortening the laser pulse is that the ablation process becomes more violent as the pulse duration is shortened. In this case photomechanical interactions, including explosive vaporization, spallation, forceful ejection and laser-induced acoustic or shock waves are generated which have the potential to cause mechanical damage a significant distance from the optical interaction zone. Hence, the quest for optimal laser parameters continues.

21.2.1 Ablation with the FEL

Wavelength effects. Traditionally, IR laser ablation of soft tissues has focused on using the strong water absorption peak around $\lambda = 2.94$ μm to heat the tissue rapidly beyond vaporization temperatures, thereby removing tissue during this explosive process. Later, the researchers begun to consider targeting various absorption bands of proteins to drive the ablative process. Edwards et al. [1] investigated several wavelengths between $\lambda = 3.0$–6.85 μm using cadaver cornea and the Mark-III FEL. It was observed that a maximum ablation yield and minimal collateral damage occurred at a wavelength of $\lambda=6.45$ μm, which corresponds to the amide-II absorption band of protein. These results were compared to those at $\lambda=2.94$ μm which had a lower ablation yield and more thermal damage and were explained by the fact that, at $\lambda=2.94$ μm, the macropulse of the FEL is not thermally confined, allowing heat to be conducted into the tissue during the laser pulse. A direct comparison between wavelengths, however, is difficult as with wavelength changes, the absorption coefficient (and thus the penetration depth) changes, which, in turn, changes the magnitude of laser-induced pressure waves. Moreover, in none of these studies was the transient dynamics in optical properties taken into account. Edwards et al. [1] originally proposed an ablation model involving a partitioning of the absorbed energy between the protein (amide-II bond) and water components of tissue. It was postulated that this energy heats the structural proteins of the tissue beyond its melting point which compromises the mechanical integrity of the tissue while the liquid–vapour transition of the heated water develops a pressure head driving the ablation. Since then, several studies have presented evidence that protein absorption does not play a significant role in the ablation process. The findings by Tribble et al. [14] suggest that the ablation process is governed by the overall absorption cross section of the sample, rather than the selective absorption of the individual components. Auerhammer et al. [30] using the FELIX found that the ablation thresholds exhibit a clear inverse proportionality to the overall corneal absorption coefficient, which was given by the weighted sum of the absorption coefficients of the individual chromophores. While some of the data presented by Auerhammer et al. [30] suggested that protein absorption plays a minimal role, no conclusive evidence on the role of amide bond absorption on the ablation process was found. Compared to measurements from Edwards et al. [1], similar ablation rates were found; however, the strong increase in ablated mass at 6.45 μm as compared to other protein absorption bands was not seen. Auerhammer noticed that at wavelengths where water is the only absorber ($\lambda=14.5$ μm), the ejected particles appeared as large chunks, while at wavelengths where protein contributes to the absorption cross section, the particles were much smaller. Also, at $\lambda=6.2$ and 6.45 μm, it was observed that the ablation plumes appeared as a fine mist during subsequent pulses when irradiated with multiple shots. This observation was attributed to the melting of the collagen matrix, in a manner similar to the

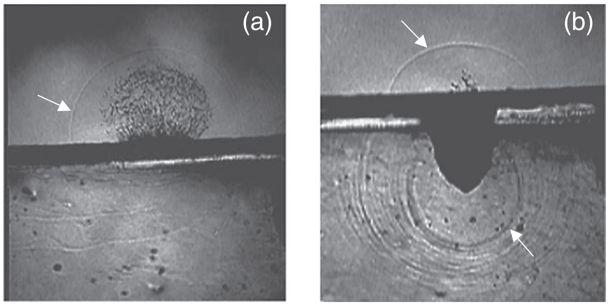


FIGURE 21.2 Pump-probe imaging of the FEL ablation of gelatin. Pulse energy=5.5 mJ pulse^{-1}, spot size=500 μm. (a) Image taken during first FEL pulse 15 μs after start of the laser pulse. Clearly visible is the ablation plume and the laser-induced pressure wave (white arrows). (b) Image taken during fifth FEL pulse delivered to the same spot, 15 μs after start of the laser pulse. Once a crater is formed, liquefied water in the crater is ablated and more efficient coupling of pressure transients into the ablation target is seen. (Figure adapted from [15].)

explanation put forth by Edwards et al. [1]. It should be noted that this observation is also consistent with observations of UV soft tissue ablation, where protein is the dominant chromophore. Here, it is commonly seen that the material ejected is composed of a fine mist [31], which suggests that protein absorption in the IR may play a role in the ablation process, as it does in the UV. An example of the Mark-III FEL-induced ablation plume (λ=6.45 μm) is shown in Figure 21.2.

Pulse structure effects. The studies discussed thus far have investigated the effect of wavelength and chromophores on the ablation process but have neglected to address the issue of the FEL pulse characteristics. The pulse structure of the Mark-III FEL places it in a unique temporal regime. The pulse consists of a ~5 μ*s* macro-pulse that is made up of a 3-GHz train of micro-pulses with a 1 ps duration. This pulse structure presents the possibility of producing a number of types of interactions, ranging from purely photothermal to possibly photomechanical effects due to the short-duration micro-pulses. The type of interaction that is likely to be seen largely depends on the relationship between the absorption properties of the tissue and the laser pulse duration. Depending on the relationship between these two parameters, the irradiation of tissue can enter certain energy confinement conditions, outlined below [32]. Thermal confinement in laser ablation occurs when the laser pulse duration (τ_p) is shorter than the time it takes heat to diffuse out of the irradiated volume. When the penetration depth is much less than the laser spot radius, the condition for thermal confinement is given by

$$\tau_{\text{th}} = \frac{\delta^2}{4\alpha}$$

where δ is the absorption depth of the laser and α is the thermal diffusivity of the tissue (α=0.15 mm^2 s^{-1} for water at 37°C). For laser pulses that are thermally confined, the temperature distribution is determined by the laser light distribution [33]. Walsh and Deutsch [34] found that by keeping the pulse duration shorter than the thermal diffusion time (τ_{th}), the thermal damage to the surrounding tissue was minimized and the depth of thermally damaged tissue increased with increasing pulse duration. It should be noted that the FEL macro-pulse of ~5 μs is thermally confined for all wavelengths except at the water absorption peak at 2.94 μm, where τ_{th}=1.7 μs. When a laser pulse is shorter than the time it takes a stress wave to propagate out of an irradiated volume (τ_{stress}), it is said to be inertially confined. The criterion for inertial (or stress) confinement is:

$$\tau_{\text{stress}} = \frac{\delta}{\sigma}$$

where δ is the absorption depth and σ is the speed of sound in tissue (σ=1500 m s^{-1}). An inertially confined laser pulse can generate large peak stresses that may contribute to the ablation mechanism and inflict damage to the surrounding tissue. Some of the photomechanical interactions that stress-confined laser pulses can elicit are the thermoelastic expansion due to the heating of the target and the recoil caused by the forced ejection of the ablated material. The ultrashort micro-pulses of the FEL (1 ps) are stress-confined at all wavelengths. Even the micro-pulse separation of 350 ps is short compared to the confinement time, since at λ=2.94 μm, τ_{stress}=670 ps. While consideration of the FEL pulse characteristics leads one to expect a photothermal or photomechanical mechanism, the ultra-short micro-pulse duration also led investigators to consider photochemical effects resulting from fast protein dynamics. Edwards et al. [35] compared the results of FEL irradiation with that of an OPA, both operating at λ=6.45 μm and containing picosecond duration micro-pulses; however, the OPA has a kilohertz repetition rate which reduces its average power by a factor of 10^6 compared to the FEL. Irradiation of ocular tissue with the OPA showed no observable effect, while the FEL ablated the tissue with clear, observable craters. This implied that picosecond-time-scale protein dynamics were not a factor in the ablation process. This led Edwards et al. [36,37] to refine the original proposed model of selective absorption by developing a heat diffusion model of tissue ablation that incorporates the chemical kinetics of protein denaturation. Edwards concluded that the differential absorption of protein denatured the protein matrix of the tissue, thereby affecting its material strength.

21.3 Towards Clinical Application of the FEL

Although the exact mechanism of FEL ablation is still subject to debate [4,14,15,38–44], a natural step in the evolution of this technology as a precise cutting tool is to move on to human clinical studies. Extensive pre-clinical studies, primarily in neurosurgery and ophthalmology, have shown that the FEL is capable of precise tissue ablation with minimal collateral damage [1,8,45,46]. While there are a significant number of FEL facilities in the world, Vanderbilt's W M Keck Free Electron Laser Center is a state-of-the-art research facility, dedicated to the development of applications of FELs to medicine and materials science and is the only FEL which integrates patient care, animal experimental facilities, physics and cell biology labs in the same building as the FEL itself [2,3]. In addition to the local user community, several research groups from around the world make regular use of the centre and its infrastructure. Promising results, especially in the field of medicine [1,4,8,39,45–51] formed the basis for a two-floor addition on top of the initial two-floor building, the top floor of which is equipped as a hospital with surgical facilities (completed in 1996). Continuous developments in the centre have demonstrated that the reliability and the performance of the Mark III FEL at Vanderbilt satisfy the requirements of a medical laser system [52]. However, before such a step can be accomplished, the issue of clinical beam delivery of the FEL beam, amenable to use in a clinical setting, must be addressed. The specific challenges can be summarized as follows:

1. the delivery system has to be able to accommodate the tunability of the laser (i.e. one or a couple of range-specific systems needs to be able to cover the tuning range of the laser (2–10 μm));
2. the delivery system has to be able to deal with the high-peak irradiances delivered by the laser ($\sim 1.0 \times 10^9$ W cm^{-2}, assuming 40 mJ macro-pulse^{-1}, 5 μs macro-pulse duration, 1 ps micro-pulse duration and a 600 μm spot size); and
3. a significant portion of the wavelength bands covered by the laser are highly absorbed by the atmosphere, thus making free beam propagation particularly lossy.

Given the infrastructure, beam delivery of the FEL involves two steps. First, the laser beam needs to be guided from the laser cavity which is located in the basement of the building to one of the ORs located four floors and approximately 100 m away. Second, the laser energy needs to be delivered by a handheld device controlled by the surgeon and focused onto the target tissue. In both parts of the system, the losses (due to absorption and reflections) must be kept to a minimum.

21.3.1 Beam Transport and Alignment

The design of a stable, reliable and user-friendly beam switching and alignment system for the FEL, which transports the beam from the laser cavity to the operation field, is a unique task. Operating in the wavelength range beyond 2.5 μm requires unconventional new beam delivery techniques due to the wide tunability of the FEL, strong atmospheric absorption at some wavelength bands (e.g. μ_a, $\mu_{a,6.45}$ $\mu_m = 0.25$ m^{-1} at a relative humidity of 50%, i.e. the energy dropped to $1/e$ of its initial value after a path length of 4 m), the high peak intensity of the micro-pulses and the long distance between the laser source and target operating field. Because of infrared absorption in the atmosphere, the beam is transported to the laboratories and to the surgical suites in an evacuated pipe system. A total of 14 mirrors direct the laser beam from the laser out-coupler into the OR. The evacuated beam transport system terminates in a ceiling-mounted tank in the OR and the laser beam is transported through a BaF_2 window that caps the evacuated beam transport system. In this tank the beam is either reduced to match the entrance aperture of an articulated mirror arm using a two-lens telescope or it is refocused using a single lens into a hollow waveguide-based, hand-held delivery probe [53–56]. Located on the same floor near the OR is a laser contingency room, where the laser parameters delivered to the operation field are controlled and monitored. A beam-switching system controls to which outlet in the OR the beam is directed. Alignment, pointing stability, reproducibility and safety concerns of the beam transport system are addressed in the design. Overall losses in the beam transport and alignment system owing to mirrors, windows and lenses amount to at least 40%.

21.3.2 Surgical Beam Delivery

We have successfully developed, implemented and tested beam delivery systems for the FEL [52,54,55,57]. These systems have to date been used in surgical applications (neurosurgery and ophthalmology) on seven patients. In brief, there are four potential beam delivery modalities:

i. Free beam delivery directed by two galvo mirrors driven by a computer system (CAST=computer-assisted surgical techniques) [57]. This system consists of a fixed assembly of the mirrors and a focusing lens above the operating table and is purged with nitrogen to reduce atmospheric absorption. The system is useful and works well in applications where surface ablation of an easy-to-reach target is intended. However, it is not feasible as a beam delivery modality in a portable unit.
ii. Modified articulated arm. The articulated arm is a commonly used device for CO_2 laser applications in medicine. We have modified this mechanical mirror device to accommodate the broad tunability of the FEL. The advantages of this system are that it is well established in the medical community and has low losses. The disadvantages are that the alignment of the arm is difficult and the device is bulky (and thus not amenable to delivery in tight and hard to reach spaces) and expensive.
iii. Hollow waveguides (HWG). These waveguides are, in essence, hollow-core glass tubes with an OD<1 mm coated on the inside with a dielectric material. Harrington *et al* have optimized the waveguides for transmission of wavelengths in the 6–7 μm part of the IR spectrum [58–62]. A clinical delivery system for neurosurgical and ophthalmological applications based on the HWG technology has been developed [54–56]. This delivery system has been used clinically for ophthalmologic applications and performed well. The advantages are the flexibility and relative ease of handling of the HWG. In addition the size of the waveguides allows for use in microsurgical instruments. Efficient light coupling into the HWG is difficult owing to the small numerical aperture equivalent of the waveguide (0.06 at $\lambda = 6.4$ μm, corresponding to an acceptance angle of 3.4°) [54,55,60–62]. Moreover, the guides are lossy and

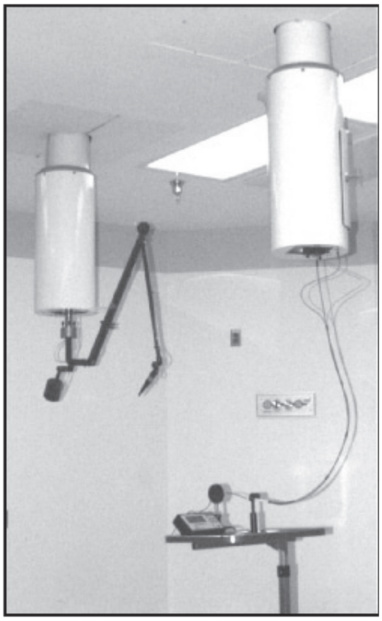


FIGURE 21.3 Picture of the two beam ports in the OR of the FEL Center: the articulated mirror arm is mounted on the wall outlet and the hollow waveguide hand-held probe is connected to the centre outlet. Note that the evacuated beam transport system terminates in the ceiling-mounted tank.

change transmission characteristics when they are bent. Similar to the articulated arm, the HWG delivery device must be purged with dry nitrogen in order to limit atmospheric absorption in the ~2 m long delivery device.

iv. Infrared transmitting fibre-optics. Figure 21.3 shows the ceiling-mounted beam delivery tanks in the OR of the Vanderbilt FEL Center. In the figure one tank is equipped with the hollow waveguide-based surgical delivery probe whilst the other is equipped with the articulated arm. There are several infrared-transmitting fibre optics materials available. Traditionally, these fibres have been brittle, lossy, toxic and/or water soluble. However, significant progress has been made in this field. Investigators at the Naval Research Laboratory are developing, improving and testing chalcogenide fibre materials for this purpose [63] while Katzir et al. has been doing the same with silver-halide fibre materials [64,65]. Data from both these collaborations show that high-irradiance infrared laser pulses can be delivered by means of specialized fibre optics that are amenable to medical applications. Damage thresholds for the FEL radiation were approximately 9×10^8 W cm^{-2} and 8×10^8 W cm^{-2} (micro-pulse peak irradiance) at 6.1 μm for chalcogenide fibres (arsenic sulphite) and silver halide fibres (AgClBr), respectively, and were found to depend on the macro-pulse repetition rate probably due to heat dissipation at the in-couple end of the fibre tip [66].

21.4 Clinical Experience

21.4.1 Neurosurgery

The first patient to be treated with the FEL (December 1999 by Copeland [67]) was a 78-year-old female with the diagnosis of a benign brain tumour, a meningioma, of at least 3 cm in size, which was easily accessible surgically (according to the FDA's IDE and the IRB protocol). Her tumour had been diagnosed 2 years earlier but it continued to grow slowly, increasing the likelihood of her symptoms becoming more severe (including seizures, headaches, weakness, ataxia, numbness and/or confusion). After the patient had opted to have her tumour removed, she was presented with the option of participating in the FEL study and informed consent was obtained. The patient was in good health and had normal neurological function. On the day of the surgery, the patient was brought from the main hospital to the FEL Center where she was prepared for surgery. The tumour was exposed using standard-of-care neurosurgical techniques, including an image-guided surgical navigation system (using pre-operative CT images). After standard exposure of the tumour, the FEL at $\lambda = 6.45$ μm was used to ablate part of the tumour roughly 5.5 mm×5.5 mm×3.0 mm in size from the outer surface of the tumour. A total of approximately 1500 pulses of 30 mJ pulse^{-1} at a spot size of 310 μm diameter were delivered. The pulse-repetition rate was 30 Hz and the laser beam was delivered to the tissue via the modified articulated arm. Pulse bursts were delivered under control of a footswitch operated by the surgeon. The remainder of the tumour was resected 'en block' and examined histologically. There were no peri-operative complications. The patient was discharged from the hospital three days after the surgery with normal neurologic condition. Immediate and three-month follow-up MRI scans showed no evidence of local brain injury or residual tumour. Several images from this first case are shown in Figure 21.4. Histologic examination of the laser crater (Figure 21.5) in the tumour showed minimal injury to adjacent cells (<62.5 μm) consistent with the earlier animal data [68,69]. Since this first case, two additional neurosurgical procedures have been performed with the FEL with similar results. In the future, more challenging resections of malignant brain tumours (and tumour margins) with the FEL or equivalent dedicated non-FEL laser systems may become a reality.

21.4.2 Clinical Experience: Ophthalmology

Shortly after the first neurosurgical procedure with the FEL, the first ophthalmological FEL procedure was performed. Based on extensive animal studies (rabbits and monkeys) [45], the FEL was hypothesized to produce efficiently an optic nerve sheath fenestration with minimal damage to the underlying axons. Normally cutting the optic nerve sheath is used to treat a condition called pseudotumour cerebri, a relatively common neurological illness among young, obese women. A build-up of cerebral-spinal fluid in the optic nerve causes blurred vision, headaches and even loss of vision. Surgery is called for when the condition does not respond to dietetic and medical

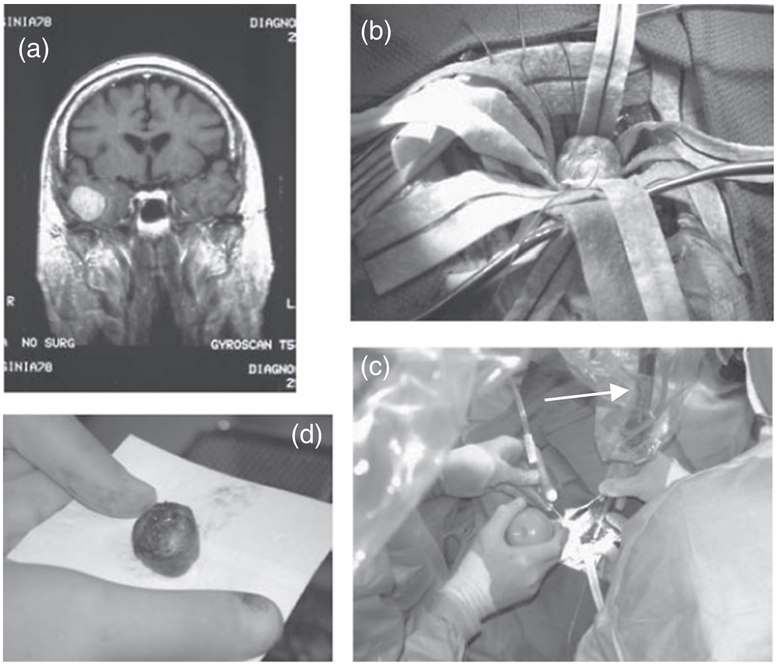


FIGURE 21.4 Images from the first clinical FEL procedure: (a) a meningioma, approximately 3 cm in size located in the right temporal lobe (pre-op T1-weighted MRI scan); (b) after surgical exposure, the tumour is readily accessible for the laser-delivery arm; (c) laser delivery via the articulated arm (arrow) ablates a divot in the tumour—the surgeon holds the handpiece and controls laser delivery via a footswitch; (d) tumour is removed with conventional means and fixed for histological analysis.

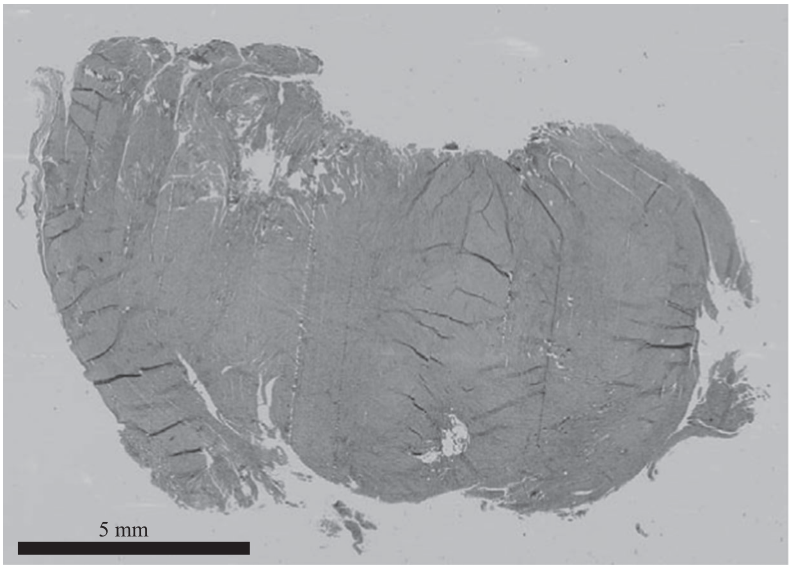


FIGURE 21.5 Histologic (H&E stained) section of the meningioma from the first FEL patient. The ablation crater is visible at the top part of the tumour. Collateral damage is small varying from undetectable to one cell width. At no point along the crater did the collateral damage exceed three cell widths (62.5 μm).

treatments. Cutting a small opening in the sheath surrounding the optic nerve relieves the pressure build-up, preventing further vision loss and, in some cases, even restoring lost vision. The FDA-approved IDE and the IRB protocol indicated initial FEL eye surgery on 10 patients (five FEL and five scissors) with end-stage traumatic glaucoma who were scheduled for enucleation. After disinsertion of the medial rectus muscle, an optic nerve sheath fenestration (2 mm diameter) was made with either the FEL (6.45 μm, 3 mJ pulse^{-1}, 325 μm spot size, 30 Hz) delivered via a hollow waveguide-based probe or with scissors. The enucleation was then completed, and the optic nerve was dissected and fixed for histological examination. Preliminary analysis shows (i) the feasibility of performing this procedure with the FEL producing adequate exposure of the optic nerve; (ii) similar histological results with the FEL and scissors and (iii) superior manoeuverability of the FEL probe compared to the scissors in this technically challenging surgical procedure [53].

21.5 Conclusions

The FEL has the potential to be used as a surgical device. Several wavelengths, unique to specific tissues, have been identified that allow precise cutting with minimal collateral damage. Nevertheless, it is highly unlikely and undesirable that the FEL will become a mainstay with widespread medical use. Instead, its role as investigational device that can be used to determine optimal laser parameters is relevant and attractive. In the near future, further developments in conventional laser technology (solid state or gas) will facilitate this approach. On the fundamental side of the interaction of the FEL with tissue, the role of pulse structure is recognized but still poorly understood. Finally, with continually improving diagnostic methods and improved spatial resolution of modalities of medical imaging, the need for therapeutic methods can take advantage of this information; treating with equivalent spatial precision will be paramount.

REFERENCES

1. Edwards G, Logan R, Copeland M, Reinisch L, Davidson J, Johnson J B, Maciunas R, Mendenhall M, Ossoff R, Tribble J, Werkhaven J and O'Day D 1994 Tissue ablation by a free-electron laser tuned to the amide II band *Nature* **371** 416–19.
2. Brau C A and Mendenhall M H 1992 The Vanderbilt free-electron laser center *Nucl. Instrum. Methods A* **319** 47.
3. Edwards G 1995 Biomedical and potential clinical applications for pulsed lasers operating near 6.45 is μm *Opt. Eng.* **34** 1524–5.
4. Bryant G L, Davidson J M, Ossoff R H, Garrett C G and Reinisch L 1998a Histologic study of oral mucosa wound healing: a comparison of a 6.0 to 6.8-micrometre pulsed laser and a carbon dioxide laser *The Laryngoscope* **108** 13–17.
5. Copeland M, Cram G P, Edwards G S, Ernst D, Gabella W and Jansen E D 2000 First human surgery with a free-electron laser *Biomedical Applications of Free Electron Lasers* ed G Edwards and R Sutherland (San Jose, CA: SPIE Biomedical Optics Symposium).
6. Izzo A D, Mackanos M A, Beckham J T and Jansen E D 2001 *In vivo* optical imaging of expression of vascular endothelial growth factor following laser incision in skin *Lasers Surg. Med.* **29** 343–50.
7. Joos K M, Shen J H, Shetlar D J and Casagrande V A 2000 Optic nerve sheath fenestration with a novel wavelength produced by the free electron laser (FEL) *Lasers Surg. Med.* **27** 191–205.
8. Peavy G M, Reinisch L, Payne J T and Venugopalan V 1999 Comparison of cortical bone ablations by using infrared laser wavelengths 2.9 to 9.2 micron *Lasers Surg. Med* **25** 421–34.
9. Robbins J B, Reinisch L and Ellis D L 2001 Optimizing free electron laser incisions at 6.45 micron with computer-assisted surgical techniques and the utilization of a heat-conducting template *Lasers Surg. Med.* **28** 162–7.
10. Sun W, Shen J H, Shetlar D J and Joos K M 2000 Endoscopic goniotomy with the free electron laser in congenital glaucoma in rabbits *J. Glaucoma* **9** 325–33.
11. Edwards G, Engh D, Kozub J and Williams R 1998 Infrared dynamics of collagen, microtubules, and water: biophysical research enabling biomedical FEL applications *Nucl. Instrum. Methods Phys. Res. B* **144** 260–4.
12. Edwards G S 1998 Physical mechanisms governing the ablation of biological tissue *Laser Ablation and Desorption* ed J C Miller and R F Haglund (San Diego, CA: Academic), 449–73.
13. Jansen E D, Pratisto H S, Uhlhorn S R and Harrison S 1999 Tissue ablation with the free electron laser: contributions of wavelength and pulse structure *Laser-Tissue Interaction X (Proc. SPIE 3601)* ed S L Jacques, G J Mueller, A Roggan and D H Sliney (Bellingham, WA: SPIE), 356–61.
14. Tribble J, Lamb D C, Reinisch L and Edwards G 1997 Dynamics of gelatin ablation due to free-electron-laser irradiation *Phys. Rev. E* **55** 7385–9.
15. Uhlhorn S R, Harrison S, Pratisto H S and Jansen E D 1999 Tissue ablation with the free-electron laser: contributions of wavelength and pulse structure *Laser-Tissue Interaction X (Proc. SPIE 3601)* ed S L Jacques (Bellingham, WA: SPIE) pp 356–61.
16. Uhlhorn S R, London R A, Makarewicz A J and Jansen E D 2000 Hydrodynamic modelling of tissue ablation with free-electron laser *Laser-Tissue Interaction XI (Proc. SPIE 3914)* ed D D Duncan, J O Hollinger and S L Jacques (Bellingham, WA: SPIE) pp 238–43.
17. Uhlhorn S R, Mongin J D, Mackanos M A and Jansen E D 2001 Effect of IR wavelength on ablation mechanics: a study of acoustic signals *Laser-Tissue Interaction XI (Proc. SPIE 4257)* ed S L Jacques (Bellingham, WA: SPIE) pp 178–83.
18. Cummings J P and Walsh J T 1993 Tissue tearing caused by pulsed laser-induced ablation pressure *Appl. Opt.* **32** 494–503.
19. Puliafito C A, Stern D, Krueger R R and Mandel E R 1987 High-speed photography of excimer laser ablation of the cornea *Arch. Ophthalmol.* **105** 1255–9.
20. Srinivasan R, Dyer P E and Braren B 1987 Far-ultraviolet laser ablation of the cornea: photoacoustic studies *Lasers Surg. Med.* **6** 514–19.
21. Venugopalan V, Nishioka N S and Mikic B B 1995 The thermodynamic response of soft biological tissues to pulsed ultraviolet laser irradiation *Biophys. J.* **69** 1259–71.
22. Venugopalan V, Nishioka N S and Mikic B B 1996 Thermodynamic response of soft biological tissues to pulsed infrared-laser irradiation *Biophys. J.* **70** 2981–93.
23. Walsh J T and Deutsch T F 1991 Measurement of Er:YAG laser ablation plume dynamics *Appl. Phys.* B **52** 217–24.
24. Buell B R and Schuller D E 1983 Comparison of tensile strength in CO_2 laser and scalpel skin incisions *Arch. Otolaryngol.* **109** 465–7.
25. Cochrane J P, Beacon J P, Creasey G H and Russell R C 1980 Wound healing after laser surgery: an experimental study *Brit. J. Surg.* **67** 740–3.
26. Hall R R 1971 The healing of tissues incised by a carbon-dioxide laser *Brit. J. Surg.* **58** 222–5.
27. Hambley R, Hebda P A, Abell E, Cohen B A and Jegasothy B V 1988 Wound healing of skin incisions produced by ultrasonically vibrating knife, scalpel, electrosurgery, and carbon dioxide laser *J. Dermatol. Surg. Oncol.* **14** 1213–17.

28. Speyer M, Joe J, Davidson J M, Ossoff R H and Reinisch L 1996 Thermal injury patterns and tensile strength of canine oral mucosa after carbon dioxide laser incisions *Laryngoscope* **106** 845–50.
29. Taylor D L, Schafer S A, Nordquist R, Payton M E, Dickey D T and Bartels K E 1997 Comparison of a high power diode laser with the Nd:YAG laser using in situ wound strength analysis of healing cutaneous incisions *Lasers Surg. Med.* **21** 248–54.
30. Auerhammer J M, Walker R, van der Meer A F C and Jean B 1999 Dynamic behaviour of photablation products of corneal tissue in the mid-IR: A study with FELIX *Appl. Phys. B* **68** 111–19.
31. Puliafito C A, Steinert R F, Deutsch T F, Hillenkamp F, Dehm E J and Adler C M 1985 Excimer laser ablation of the cornea and lens *Experiment. studies Ophthalmol.* **92** 741–8.
32. Jacques S L 1992 Laser-tissue interactions. Photochemical, photothermal, and photomechanical *Surg. Clin. North Am.* **72** 531–58.
33. Welch A J and Gemert M J C van (ed) 1995 *Optical-Thermal Response of Laser-Irradiated Tissue* (New York: Plenum), 709–63.
34. Walsh J T Jr and Deutsch T F 1988 Pulsed CO_2 laser tissue ablation: measurement of the ablation rate *Lasers Surg. Med.* **8** 264–75.
35. Edwards G S, Hutson M S, Hauger S, Kozub J, Shen J, Shieh C, Topadze K and Joos K 2002 Comparison of OPA and Mark-III FEL for tissue ablation at 6.45 μm *Commercial and Biomedical Applications of Ultrafast and Free-Electron Lasers (Proc. SPIE 4633)* ed G S Edwards, J Neev, A Ostendorf and J C Sutherland, 194–200.
36. Edwards G S, Hutson H S and Hauger S 2002 Heat diffusion and chemical kinetics in Mark-III FEL tissue ablation *Commercial and Biomedical Applications of Ultrafast and Free-Electron Lasers (Proc. SPIE 4633)* ed G S Edwards, J Neev, A Ostendorf and J C Sutherland, 184–93.
37. Hutson M S, Hauger S and Edwards G S 2002 Thermal diffusion and chemical kinetics in laminar biomaterial due to heating by a free-electron laser *Phys. Rev. E* **65** 1–6.
38. Awazu K, Nagai A, Tomimasu T and Aizawa K 1998 An innovative method to selectively ablate cholesterol esters from an arteriosclerotic region of the arterial wall using a free-electron laser *Nucl. Instrum. Methods Phys. Res. B* **144** 255–28.
39. Bryant G L, Lano C F J, Reinisch L, Ossoff R H and Werkhaven J A 1998b Ablation of Teflon granuloma with the free-electron laser emitting in the eight- to nine-micron range *Ann. Otol., Rhinol., Laryngol.* **107** 269–74.
40. Ellis D L, Weisberg N K, Chen J S, Stricklin G P and Reinisch L 1999 Free electron laser infrared wavelength specificity for cutaneous contraction *Lasers Surg. Med* **25** 1–7.
41. Fortune D S, Huang S, Bryant G L, Garret C and Reinisch L 1998 Effect of laser-pulse structure and wavelength on wound healing *Lasers in Surgery: Advanced Characterization, Therapeutics, and Systems VIII (Proc. SPIE 3245)* ed R R Anderson (Bellingham, WA: SPIE), 191–9.
42. Fowler W C, Rose J G, Chang D H and Proia A D 1999 The free electron laser: a system capable of determining the gold standard in laser vision correction *Nucl. Instrum. Methods Phys. Res. A* **429** 497–501.
43. Lamb D C, Doukas A G, Flotte T J, Ossoff R H, Reinisch L and Tribble J 1995 Free-electron laser as the ideal stress-wave generator *Laser-Tissue Interaction VI (Proc. SPIE 2391)* ed S L Jacques (Bellingham, WA: SPIE), 192–201.
44. Lamb D C, Tribble J, Doukas A G, Flotte T J, Ossoff R H and Reinisch L 1999 Custom designed acoustic pulses *J. Biomedical Opt.* **4** 217–23.
45. Joos K M, Edwards G S, Shen J-H, Shetlar D, Robinson R and O'Day D 1996 Free electron laser (FEL) laser-tissue interaction with human cornea and optic nerve *Ophthalmic Technologies VI (Proc. SPIE 2673)* ed J-M Parel, K M Joos and P O Rol (Bellingham, WA: SPIE), 89–92.
46. Shetlar D, Joos K, Shen J and Robinson R 1997 Endoscopic goniotomy with the free electron laser *Invest. Ophtalmol. Vis. Sci. (Suppl.)* **38** 169.
47. Harris D M, Reinisch L, Edwards G S, Yessik M J, Ashrafi S and Santos-Sacchi J 1996 Midinfrared ablation of dentin with the Vanderbilt FEL *Lasers in Dentistry II (Proc. SPIE 2672)* ed H A Wigdor, J D Featherstone, J M White and J Neev (Bellingham, WA: SPIE), 165–75.
48. Kelly J H and Qadir R H 1996 Laser decompression of the facial nerve *Otolaryngol. Clinics North America* **29** 1049–61.
49. Lano C J, Reinisch L, Ossoff R, Garrett C, Kuo T, Bryant G J and Werkhaven J 1999 Ablation of Teflon granulomas in the canine larynx with the free-electron laser *Ann. Otol. Rhinol. Laryngol.* **108** 17–23.
50. Reinisch L and Ossoff R H 1993 Plasma ablation of hard tissue by the free-electron laser *Free-Electron Laser Spectroscopy in Biology, Medicine, and Materials Science (Proc. SPIE 1854)* ed H A Schwettman, 145–53.
51. Shen J-H, Casagrande V A, Joos K M, Shetlar D J, Robinson R D, Head W S, Mavity-Hudson J A and Nunnally A H 1999 Acute optic nerve sheath fenestration with the free-electron laser *Ophthalmic Technologies IX (Proc. SPIE 3591)* ed P O Rol, K M Joos, F Manns, B E Stuck and M Belkin (Bellingham, WA: SPIE), 235–40.
52. Edwards G, Evertson D, Gabella W, R R G, King T, Kozub J, Mendenhall M, Shen J, Shores R, Storms S and Traeger R 1996 Free-electron lasers: performance, reliability, and beam delivery *IEEE J. Selected Topics Quant. Elect.* **2** 810–17.
53. Joos K M, Mawn L A, Shen J H, Jansen E D and Casagrande V A 2002 Acute optic nerve sheath fenestration in humans using the free electron laser (FEL): a case report *Ophthalmic Technologies XII (Proc. SPIE 4611)* ed F Manns, P G Soederberg and A Ho (Bellingham, WA: SPIE), 81–5.
54. Pratisto H S, Uhlhorn S R, Copeland M and Jansen E D 1999 Clinical beam delivery of the Vanderbilt FEL: design and performance of a hollow waveguide-based handheld probe for neurosurgery *Specialty Fiber Optics for Medical Applications (Proc. SPIE 3596)* ed A Katzir and J A Harrington, 55–61.
55. Pratisto H S, Uhlhorn S R and Jansen E D 2000 Beam delivery of the Vanderbilt free electron laser with hollow wave guides: effect on temporal and spatial pulse propagation *Fiber and Integrated Optics* **20** 83–94.
56. Shen J H, Harrington J A, Edwards G S and Joos K M 2001 Hollow-waveguide delivery of an infrared free-electron laser for microsurgical applications *Appl. Opt.* **40** 583–7.

57. Reinisch L, Mendenhall M, Charous S and Ossoff R 1994 Computer-assisted surgical techniques using the Vanderbilt free electron laser *Laryngoscope* **104** 1323–9.
58. Gannot I, Inberg A, Oksman M, Waynant R and Croitoru N 1996 Current status of flexible waveguides for IR laser radiation transmission *IEEE J. Selected Topics Quant. Elect.* **2** 880–9.
59. Gannot I, Waynant R, Inberg A and Croitoru N 1997 Broadband flexible waveguides for free-electron laser radiation *Appl. Opt.* **36** 6289–93.
60. Matsuura Y, Abel T and Harrington J A 1995 Optical properties of small-bore hollow glass waveguides *Appl. Opt.* **34** 6842–7.
61. Matsuura Y, Matsuura K and Harrington J A 1996a Power delivery of free electron laser light by hollow glass waveguides *Appl. Opt.* **35** 5395–7.
62. Matsuura Y, Rabii C D, Matsuura K and Harrington J A 1996b Low order multimode generation in hollow glass waveguides *Electron. Lett.* **32** 1096–8.
63. Sanghera J S, Shaw B L, Busse L E, Talley D and Aggarwal I D 1999 Infrared-transmitting fiber optics for biomedical applications *Specialty Fiber for Medical Applications (Proc. SPIE 3596)* ed J A Harrington and A Katzir, 178–87.
64. Moser F, Bunimovich D, DeRowe A, Eyal O, German A and Katzir A 1996 Medical applications of infrared transmitting silver halide fibers *IEEE J. Select Topics Quant. Elect.* **2** 872–9.
65. Nagli L, Bunimovich D, Shmilevich A, Kristianpoller N and Katzir A 1993 Optical properties of mixed silver halide crystals and fibers *J. Appl. Phys.* **7** 5737–41.
66. Mackanos M A, Shaw B L, Aggarwal I D, Katzir A and Jansen E D 2003 Delivery of mid-infrared (6–7 um) laser radiation in a liquid environment using infrared-transmitting optical fibers *J. Biomed. Opt.* submitted.
67. Edwards G S et al FEL-based biophysical and biomedical instrumentation *Rev. Sci. Instrum.* Submitted.
68. Copeland M L, Maciunas R J and Edwards G S 1998 Use of the Free-Electron Laser for metastatic brain tumours *Neurosurgical Topics: Advanced Techniques in Central Nervous System Metastases* ed R J Maciunas (Rolling Meadows, IL: American Association of Neurologic Surgeons), 113–21.
69. Cram G P and Copeland M L 1998 Hemostatic properties of the free electron laser *Nucl. Instrum. Methods Phys. Res. B* **144** 256–9.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

22

Medical Diagnostics

Brian C. Wilson

CONTENTS

22.1 Introduction 291
22.2 *In Vivo* Methods and Applications 292
  22.2.1 White-Light (Reflectance) Imaging 292
  22.2.2 Diffuse Optical Spectroscopy and Spectral Imaging 293
  22.2.3 Elastic Scattering Spectroscopy 298
  22.2.4 Optical Coherence Tomography 298
  22.2.5 Confocal Imaging 301
  22.2.6 Fluorescence Spectroscopy and Imaging 302
  22.2.7 Raman Spectroscopy and Imaging 303
    22.2.7.1 Photoacoustic Imaging 305
  22.2.9 Other Methods 306
  22.2.10 Optical Imaging in Pathology 307
22.3 Comparison of Techniques and General Future Trends 307
Bibliography 308

22.1 Introduction

The use of light for medical diagnostics exploits the changes in one or more measurable properties of ultraviolet (UV), visible (vis) or near-infrared (NIR) photons that are transmitted through or reflected from tissue due to interaction with tissue structures and molecules. There are many different techniques under development for optical diagnostics. In this chapter, only those that are or potentially may be used clinically *in vivo*, i.e. non-invasively or minimally invasively in patients, will be considered. The next section will present the underlying physical principles for each technique, show examples of clinical applications and illustrate the instrumentation involved. In each case, future developments will also be indicated. The final section will point to major challenges and overall trends in this field.

Predominantly, clinical applications involve the use of vis/NIR light, since this has high molecular specificity and also allows adequate penetration in tissue for non-invasive measurements. UV is sometimes used as an excitation source (e.g. for fluorescence), although this has very limited tissue penetration and raises a concern due to its potential mutagenicity. Beyond about 1300 nm, the absorption of water becomes too high to allow useful measurements except in a few special cases.

It is important to recognize that optical methods must be considered in the landscape of many competing diagnostic techniques: X-ray imaging, magnetic resonance imaging/spectroscopy, radionuclide and ultrasound imaging. With few exceptions, optical methods are relative newcomers, and since many of the key enabling technologies, including light sources, fibre optics and detectors, are themselves still evolving rapidly, the field is far from mature. Hence, some of the techniques and applications presented here are not yet at the clinical stage, while others may have a transient existence as useful clinical techniques. Not only must these optical methods provide information that is unique or complementary to other diagnostic methods, but the instrumentation must also be safe, as well as cost-effective and user-friendly in the intended clinical setting.

The underlying impetus for the development of optical diagnostics derives from two fundamental characteristics of light–tissue interactions.

1. UV-vis-NIR photon energies are comparable to the electronic or vibrational energy levels of molecules (~eV), giving the potential to probe molecular states. This contrasts with, say, X-rays where the interactions are at the atomic level with no sensitivity to molecular structure, or to ultrasound where the diagnostic information comes from the mechanical properties of tissues. This is important also in the development of therapeutic applications of light and in the non-clinical applications in biology and biotechnology.
2. Optical wavelengths are on the order of cellular structures (~µm), so that it is possible to achieve

> spatial resolution (as exploited, for example, in optical microscopy) that is much superior to that of any other *in vivo* diagnostic modality.

A less obvious third advantage is that it is possible to use optical 'labelling' to provide additional molecular, physiological or even genetic information that is not easily accessed by other means. Thus, for example, it is possible to conjugate fluorescent dyes to antibodies or DNA fragments and to follow these *in vivo* by fluorescence imaging or spectroscopy, as the biologically active agents bind to specific cellular targets. While radioisotope labelling can provide analogous capability, the range of possible optical compounds is much larger and there is a high degree of flexibility in the labelling chemistry. Furthermore, optical spectroscopy or spectral imaging enables a high level of information multiplexing, i.e. using a 'cocktail' of differently targeted biomarkers, each with a unique optical signature, for simultaneous detection or imaging of multiple molecular features of cells or tissues. This will become increasingly important with knowledge of the complex molecular/genetic basis of disease and the need to assess this in individual patients in order to 'personalize' treatments.

Optical diagnostic technologies have further clinical advantage that, especially using fibre-optic light delivery and collection, the instrumentation can often be made compatible with other diagnostic and therapeutic systems. Examples of this are optical monitoring of brain function or optical imaging while a patient is inside an MRI unit, and the combining of fibre-optic fluorescence or Raman spectroscopy probes with ultrasound devices during surgery.

Finally, with photonic components—sources, optical elements, detectors, etc.—becoming more and more miniaturized and integrated into consumer electronic devices such as cell phones, and being relatively low cost and user-friendly, light-based diagnostic products are uniquely positioned to address a spectrum of point-of-care diagnostics applications, including in the developing world: for example, hand-held optical imaging devices for disease detection in remote and under-served communities, with internet communication to expert clinicians in major urban hospitals.

Despite these strengths, it is important to recognize that optical methods suffer from fundamental limitations that do not apply or are less restrictive with other techniques. In particular, the optical absorption of tissue is very high even the so-called 'optical window' around 700–1100 nm and strongly attenuates the signal, while the high (elastic) scattering 'smears' the spatial information. These two factors combine to create a severe trade-off between spatial resolution and depth in tissue. At one extreme, using red/NIR light does allow optical measurements through up to 5–10 cm or more of tissue but the spatial resolution is degraded to ~cm. Conversely, it is possible to achieve micron-resolution images *in vivo* but only to a depth of a few hundred microns below the accessible tissue surface. The optimum 'sweet spot' in this trade-off is highly dependent on the particular application.

These fundamental advantages and limitations markedly influence the clinical and technological options in optical diagnostics. A further difference between optical and non-optical *in vivo* diagnostic modalities is that there are a large number of distinctly different optical instruments, often designed for a single or only a few clinical applications. This contrasts with radiological imaging, where there are only a few classes of imaging systems that tend to be generic in their applications: e.g. MRI and CT systems can be used to image any organ in any part of the body. Partly, this difference is because the performance of optical instruments needs to be tailored to their specific purpose and in part because optical technologies can be relatively inexpensive so that it is often more cost-effective to have application-specific instruments rather than building higher cost generic platform systems with multiple capabilities.

22.2 *In Vivo* Methods and Applications

As illustrated in Figure 22.1, a light beam incident on tissues undergoes several interactions, with the resulting light exiting the tissue being detected and providing diagnostic information. In some devices, interstitial fibre-optic probes may be used to deliver and detect the light at a localized region within a tissue mass.

22.2.1 White-Light (Reflectance) Imaging

The original method of optical diagnosis is, of course, simply to look at the patient or at a specific accessible organ (e.g. the eye or the tongue). The diagnosis is then based on the subjective spectral and spatial appearance. The capability of the physician's vision may be extended by optical instruments to increase the spectral range and/or sensitivity or to provide access to inner structures, e.g. in ophthalmology to examine the retina (Figure 22.2a) and in white-light endoscopy (Figure 22.2b) to examine the lining of hollow organs such as the bronchus, gastrointestinal tract, bladder or cervix. Endoscopy is particularly interesting in terms of optical technology development. Prior to the development of high-performance optical fibres, hollow tubes could be used to look into body cavities, with some advances to improve brightness and image quality through the use of lenses and cameras. Optic fibres and coherent optical fibre bundles then allowed efficient delivery of illumination light and remote, real-time (video rate) viewing of hollow organs, with high brightness and spatial resolution. In the last few years, the fibre-optic endoscope, in which

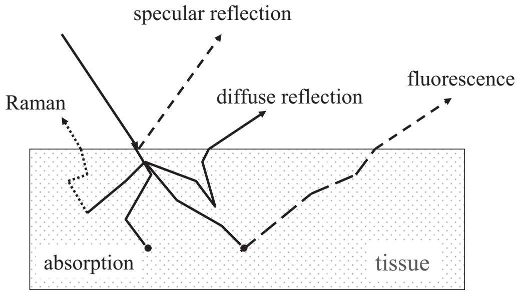


FIGURE 22.1 Schematic of the main light–tissue interactions, leading to photons exiting the tissue volume which may be detected and used to extract information on the structural and/or biochemical properties of the tissue for medical diagnostics.

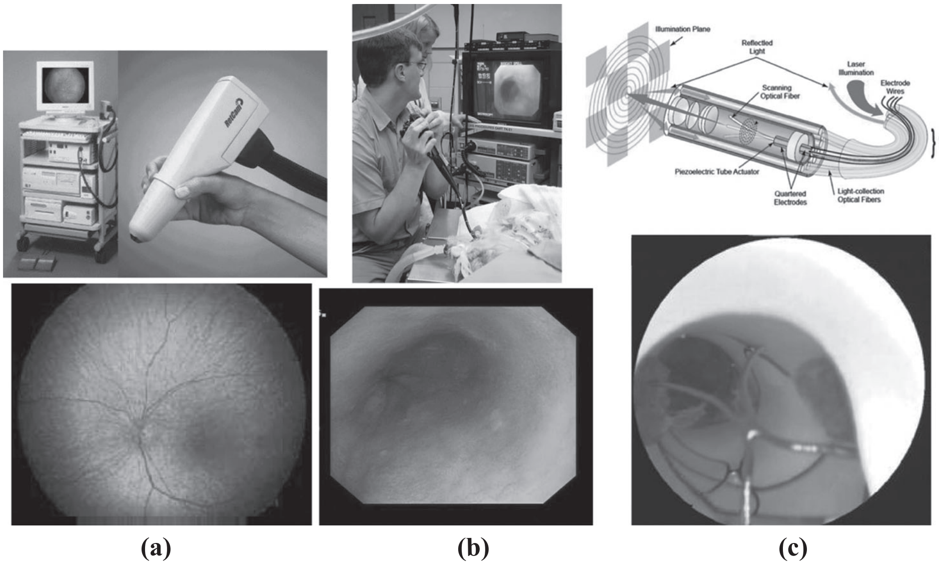


FIGURE 22.2 Examples of white-light imaging techniques: (a) wide-field retinal scope and corresponding image showing blood vessels and other structures; (b) an endoscope and frame taken from a video recording image inside the esophagus; (c) schematic of the front-end of a scanning-fibre endoscope and image within an artery showing deployment of a metal stent used to maintain blood flow in otherwise blocked vessel.

the (red-green-blue) camera is outside the body, has been replaced for many applications by video endoscopes in which a miniature RGB camera is placed at the working end of the endoscope within the patient and the image is then transferred electronically for viewing, resulting in higher image resolution. Various contrast agents such as absorbing dyes may be used to enhance the features of abnormal structures, which is of particular value in combination with high-magnification endoscopy to visualize fine surface features.

A more recent development has been scanning-fibre endoscopy (SFE) in which the image is formed by projecting a focused spot of light onto the tissue from a rapidly scanned lens-tipped single fibre, as illustrated in Figure 22.2c. Scanning is achieved using either a piezoelectric driver or MEMS device. Adjacent fibres, alongside or in a ring surrounding the source fibre, collect the light point by point as the focus is scanned across the tissue surface and transmitted to a suitable photodetector. The tip is scanned at around 10^5 Hz to generate real-time images. The advantages of SFE are that the endoscope diameter can be made very small (<~ 1 mm), allowing high-resolution imaging even within arteries, as shown in Figure 22.2c; unique to this technology, the image formation is independent of the specific light–tissue interaction mechanism, so that RGB, fluorescence, Raman and other detection schemes including non-linear techniques can be used as long as the signal-to-noise is high enough; the probes are relatively low cost or even disposable; image-guided phototherapies should be possible by switching from delivering diagnostic powers and wavelengths to high-power (laser) sources. The potential limitations are that the field-of-view depends strongly on the probe-to-tissue distance, and a separate device is required for image-guided endoscopic biopsy.

Relatively little objective or quantitative image analysis is currently in routine clinical use in endoscopy, although this is becoming more feasible with the advent of digital image capture and high-speed on-line computation. Hyperspectral imaging, in which images are captured at many separate wavelengths, is a possible future trend in endoscopy, as well as the various other modalities discussed below (fluorescence, Raman, OCT, non-linear, etc.)

22.2.2 Diffuse Optical Spectroscopy and Spectral Imaging

The dominant processes in light–tissue interactions, and the basis of techniques such as white-light endoscopy, are elastic scatter and absorption. Elastic scattering, i.e. scattering without energy loss, is very high in tissues at the vis/NIR wavelengths that are most commonly used diagnostically: the scattering coefficient (μ_s) due to micro-inhomogeneities in refractive index is typically tens to hundreds of cm^{-1}. Elastic (Mie-like)

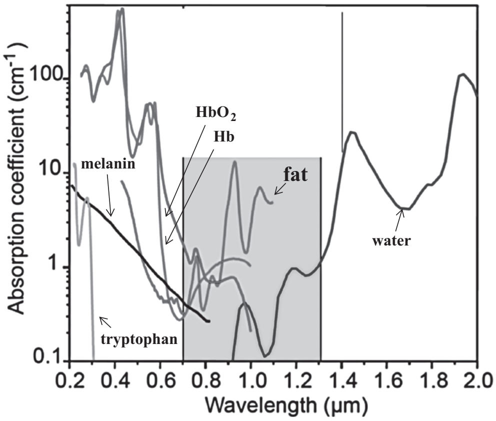


FIGURE 22.3 Some of the main light-absorbing molecules found in tissues at normal physiological concentrations. The shaded wavelength region represents the 'optical window' where the penetration of light is the greatest.

scattering is usually forward peaked with an anisotropy (mean cosine of the scattering angle), $g > \sim 0.9$, resulting in an effective (transport) scattering coefficient, $\mu'_s = \mu_s(1-g) \sim 1\text{–}10\ \text{cm}^{-1}$. This is the scattering coefficient for equivalent isotropic scattering and applies to measurements in which the angular dependence is lost through multiple scattering. Thus, light is rapidly diffused in tissue, affecting the spatial and temporal distribution of transmitted or reflected (backscattered) photons. The elastic scattering coefficient has a simple, roughly $1/\lambda$ form throughout the spectral range.

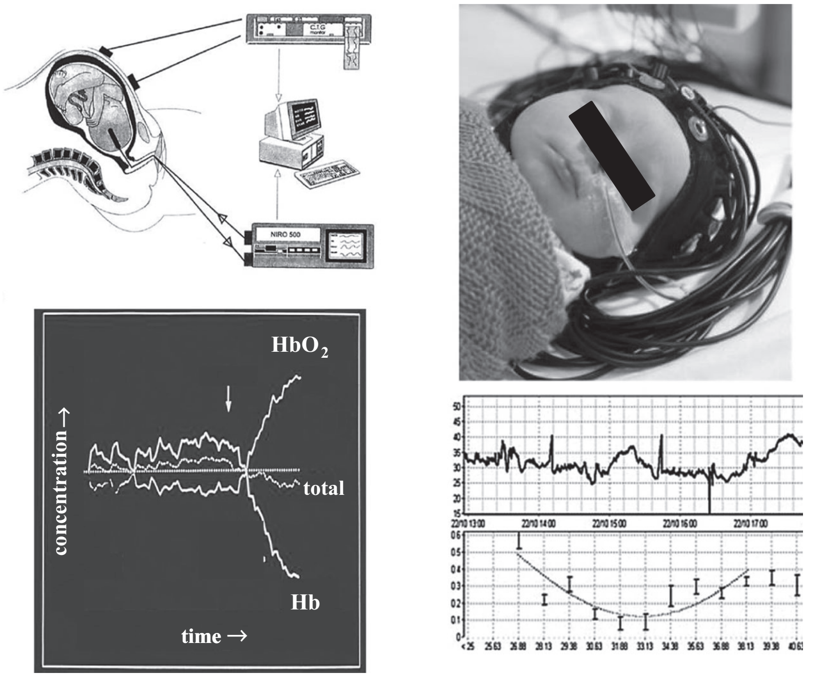
The probability of absorption of photons along their tortuous multiple scattering paths is given by the absorption coefficient, μ_a, which depends on the molecular composition. Major naturally occurring (endogenous) chromophores in the vis-NIR range include haemoglobin (both in the oxygenated, HbO_2, and de-oxygenated, Hb, forms), melanin, water, lipids and various pigments, the absorption spectra of which are shown in Figure 22.3. Other molecules such as amino acids absorb strongly in the UV, as in the example of tryptophan. The absorption coefficient of soft tissues at vis/NIR wavelengths is typically ~ 0.1–10 cm^{-1}.

In techniques such as endoscopy, the white-light image is formed from the diffusely reflected light. This is then dominated by light scattered from the surface layers of the tissue. The effective interrogation depth increases with wavelength, from blue to red. Sub-surface blood vessels can be visualized, as well as alterations in the tissue surface/sub-surface structure.

Measuring the spectrum of diffuse light from bulk tissue *in vivo* opens up many different applications, based on specific tissue absorption and/or scattering properties of the tissue. The corresponding technologies can be as simple as a white-light source, light delivery/collection optics (including fibre optics) and a single-element photodetector with selected filters, while complex imaging instruments with multiple sources and detectors and/or high-performance spectrographs may be required for some applications. The key information lies mostly in the absorption features, but there are applications where the scattering also is informative (see below). A major challenge is to separate the absorption and elastic scattering contributions from the overall measured reflectance or transmittance measurements. This requires acquisition of additional information on the spatial and/or temporal distribution of the detected light.

A common application that utilizes objective spectral measurements is to monitor the blood content ($Hb+HbO_2$) and/or oxygen saturation ($HbO_2/[Hb+HbO_2]$) of tissues, especially in the brain for continuous non-invasive monitoring of premature infants or even during birth (Figure 22.4a). Multi-source, multi-detector systems have been developed for rapid mapping of neurological activity, complementing high-resolution but slow functional magnetic resonance imaging (Figure 22.4b). Simpler, hand-held instruments have also been used for emergency monitoring in head trauma to distinguish between haemorrhage and stroke. Assessing the viability of tissue, e.g. following organ transplant or skin grafting for burns, by identifying characteristic spectral signature of HbO_2 is another application.

Other specific chromophores may also be monitored optically *in vivo*, for example, to detect circulating bilirubin in neonates with jaundice through the sharp spectral peak of the bilirubin at



[redacted]

(a) (b)

FIGURE 22.4 Monitoring haemoglobin and oxyhaemoglobin in the brain based on spectrally resolved NIR diffuse reflectance or transmittance: (a) during birth with a probe placed on the scalp and corresponding time race of brain haemoglobin and oxyhaemoglobin (arrow indicates the time at which the head is delivered and the resulting jump in brain oxygenation) (*Courtesy D. Delpy, UK*); (b) continuous monitoring in premature baby, with the traces showing the arterial (upper) and brain (lower) signals. (*Courtesy T. Austin, UK.*)

~460 nm, using a simple contact probe connected to a photodetector with appropriate filters. Monitoring of therapeutic drugs is another option and this has been developed mainly in association with photodynamic therapy to monitor the photosensitizer uptake and pharmacokinetics in individual patients to determine the optimum timing and dose of light. This is often done using fluorescence (see below), which has higher sensitivity.

Another potential application of diffuse reflectance/transmittance spectroscopy *in vivo* has been to measure glucose levels in tissue, as a non-invasive alternative to blood sampling in diabetics. In principle, this could be very simple using, say, the spectrum of light diffusely transmitted through a finger. However, the glucose absorbance is very low at physiologically relevant concentrations, which places severe demands on the stability and dynamic range of the instrument, given that it must also be cost-effective for widespread adoption, particularly for home care. The optical absorption spectrum of glucose in the red/NIR wavelength range is also rather broad and unstructured, and the measured spectra are sensitive to changes in water and other blood analytes. The use of polarimetric techniques for this application has also been pursued, based on glucose being chirally active: in this case, the challenge is also one of low signal-to-background.

A further example of diffuse (transmittance) spectroscopy is monitoring changes in the breast that may be associated with breast cancer risk. The instrumentation is straightforward, as seen in Figure 22.5: low-intensity white light is launched sequentially into the breast at several points (usually 4) to sample the entire breast volume and the corresponding transmitted spectra are detected on the opposite side. The challenge lies in the spectral analysis and the correlation of subtle changes (due to altered water, fat and, possibly, blood content of the breast) with structural and/or physiologic alterations that are associated with cancer risk.

In the above methods, the focus is on obtaining information on the intrinsic (endogenous) absorption properties of the tissue, since this is linked to the specific important chromophores. The presence of elastic scatter is essential to enable measurements to be made non-invasively but it also makes the extraction of the purely absorption component difficult (see below). An example that illustrates the potential information that lies also in the scattering characteristics *per se* is the use of reflectance spectroscopy over a wide UV-vis-NIR spectral range. Using instrumentation similar to that in Figure 22.5, but operating in reflectance geometry, the spectra from the tissue (Figure 22.6) clearly show absorption features, particularly

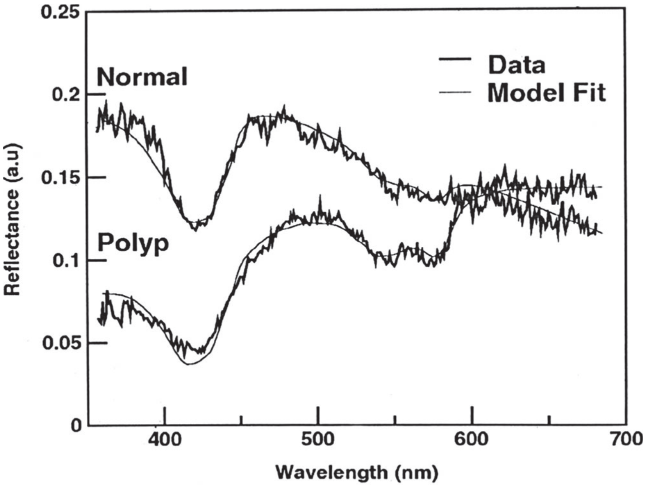


FIGURE 22.6 Example of fibre-optic based reflectance spectroscopy for tumour detection (evaluating a polyp in the colon). (*After I. Bigio, USA.*)

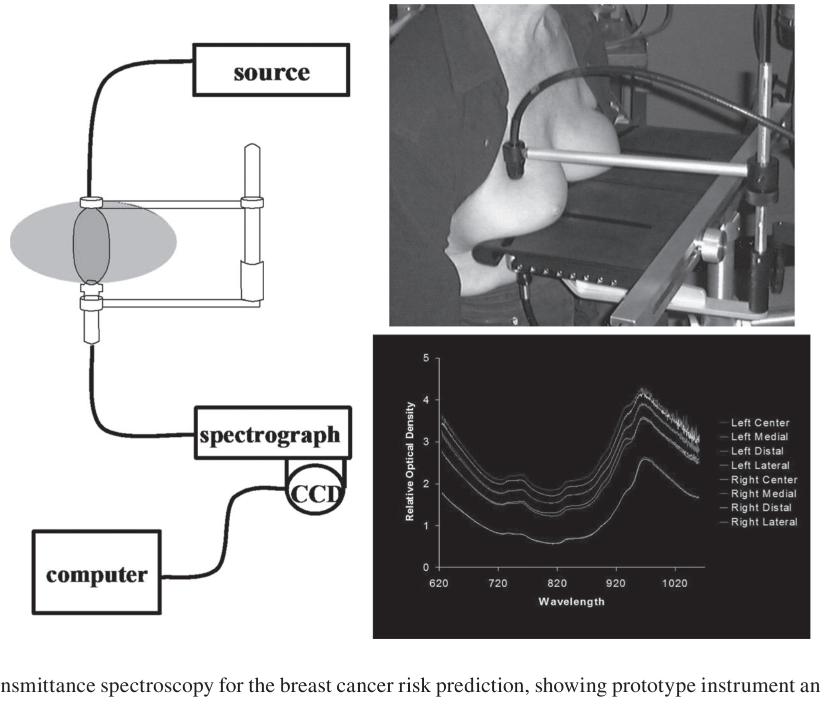
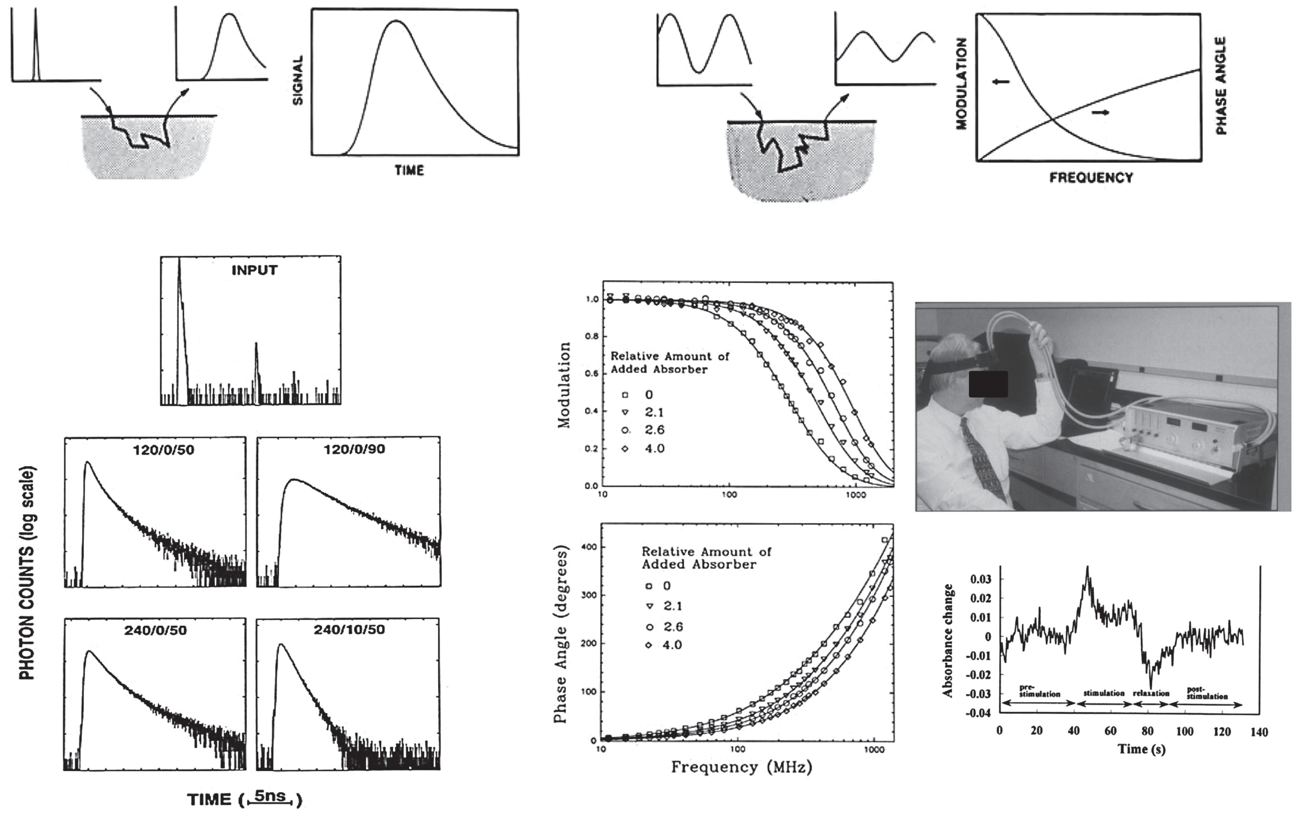


FIGURE 22.5 Diffuse transmittance spectroscopy for the breast cancer risk prediction, showing prototype instrument and typical spectra. (*Courtesy L. Lilge, Canada.*)

of haemoglobin (e.g. the double peak in the lower spectrum in the 500–600 nm range). However, detailed analysis has also shown that there can be significant differences between normal and abnormal tissues also in features that depend on altered scattering characteristics (e.g. the slope of the spectrum in the 370–420 nm range).

Implicit in these methods is the fact that the high elastic scattering means that the effective optical pathlength, i.e. the total distance travelled by a photon propagating through tissue and undergoing multiple scattering events, is typically 3–5 times longer than the physical distance between the source and the detector. Since this factor is not known *a priori*, CW diffuse spectroscopy or imaging usually yields only relative absorption measurements. A significant breakthrough was made in the late 1980s with the recognition that measuring the propagation time of photons through the tissue directly yields the optical pathlength, *l*. The average photon time-of-flight $<t>$ is simply l/c, where c is the speed of light in the tissue. For a typical average refractive index of the soft tissues of $n=1.4$, $c=c_o/n \cong 2.2\ 10^8\ \mathrm{ms}^{-1}$ or, in more practical units, approximately 20 cm per ns or 2 mm per ps. Thus, if an input laser pulse (short compared to $<t>$) is used, the detected light will be spread in time as individual photons arrive after following different tortuous paths through the tissue: examples are shown in Figure 22.7a. Analytic solutions governing this behaviour have been developed based on Diffusion Theory. Fitting these to the time curves, including appropriate boundary conditions, allows μ_a and μ’s to be determined. This is the basis for many applications of time-resolved diffuse reflectance and transmittance spectroscopy *in vivo*, including for example absolute Hb and HbO_2 concentration measurements. For more complex geometries and inhomogeneous tissues, Monte Carlo computer simulation of photon transport is often used.

In terms of instrumentation, this direct time-domain technique requires typically a laser source of a few ps pulse width. In order to obtain high average power and a high pulse repetition rate, a mode-locked, frequency-doubled Nd:YAG-pumped dye laser can be used. However, these lasers are expensive, complex and large, so they are not generally suitable for clinical use. Picosecond diode lasers are an alternative, being relatively inexpensive and compact, but have limited output power and pulse frequency and are of fixed wavelength. In order to obtain the full time curve, a sensitive photomultiplier tube can be used, operating in single photon-counting mode. With adequate laser source power, a streak camera can also be used and provides faster data collection, although the limited dynamic range can be problematic. For some applications, for example, to correct for the unknown optical pathlength through the tissue, it is not necessary to measure the full-time spectrum, thereby simplifying the detection hardware.



[redacted]

(a) (b)

FIGURE 22.7 Time-resolved light propagation in tissue: (a) temporal spread of photons detected at the surface following a ps input light pulse in media of different optical properties. The numbers on each graph are the tissue (reduced) scattering coefficient/ absorption coefficient/source-detector fibre separation on the surface; (b) analogous measurements in the frequency domain, together with an example of a clinical instrument for monitoring brain activity as a function of time.

An alternative to direct time-of-flight measurements is to use the frequency domain, as illustrated in Figure 22.7b. In this, a CW diode laser or LED source is intensity modulated at a high frequency (~10^2–10^3 MHz). The diffuse light from the tissue is then collected with a phase-sensitive detector referenced to the incident light signal and is shifted in phase because of the average time delay of the detected photons. This phase shift depends on the tissue optical properties and modulation frequency. Again, diffusion-theory equations describing the full in-phase and quadrature signals have been derived to describe this behaviour. In order to acquire the full time-dependent information, the phase shift and demodulation must be measured for all frequencies and it may not be possible to extract the optical absorption and scattering coefficients from measurements at only a single frequency. Although conceptually more complex than the direct time domain, frequency-domain measurements are technologically simpler. Conventional CW diode laser sources and off-the-shelf detectors and electronics may be used, leading to relatively compact and inexpensive instruments. Since the signal is sampled continuously, FD instruments have the great advantage of rapid response time to track physiological changes, as seen also in this figure.

More generally, diffuse optical imaging is the application of one of the above diffuse spectroscopic techniques to map or image the spatial distribution of optical absorption and scattering in tissues. Again, there are several possible approaches to address (in part) the loss of spatial resolution and contrast due to multiple scattering in imaging through thick tissue volumes, such as the breast or the brain. These fall into the categories of

a. local diffuse reflectance mapping,
b. projection imaging, in which an image is formed from the transmitted light at a single projection angle through the tissue and
c. reconstruction (tomographic) imaging, in which the diffusely transmitted light is measured at many angular positions around the tissue and cross-sectional image(s) are calculated from these measurements, as illustrated in Figure 22.8.

There is no radiological imaging analogy to (a), while (b) is analogous to conventional planar X-radiography and (c) is conceptually similar to X-ray computed tomography. However, compared to X-ray imaging, the scattering is huge and there are a negligible number of unscattered ('ballistic') photons in the transmitted light beam. At each measurement position, different optical measurement techniques may be used: CW reflectance of transmittance; full time-of-flight measurements, time-dependent diffuse reflectance or transmittance; mean time-of-flight, or phase shift and modulation at one or more modulation frequencies.

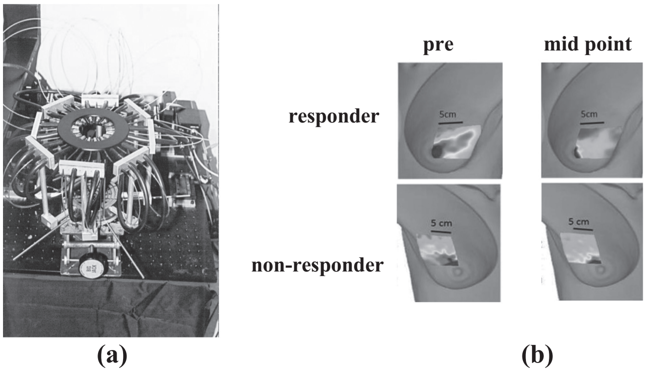


FIGURE 22.8 (a) Optomechanical configuration of a tomographic DOT system for breast imaging, showing the ring of fibre-optic light sources and detectors around the breast, which is positioned within the central opening; (b) example using this technique to monitor the response of breast cancer patients at midway through chemotherapy compared to pre-treatment baseline images in a patient whose tumour responds to treatment and a non-responding tumour in a different patient. (*Adapted from ACRIN 6691 clinical trial report.*)

As illustrated in Figure 22.8b, the presence of an optical inhomogeneity in a tissue volume (i.e. a region, such as a tumour, of higher or lower absorption and/or scattering compared with the surrounding tissue) alters the spatial and temporal light distribution. The challenge is to detect the differential signal on the large background of diffusely scattered light and, in tomographic reconstruction (diffuse optical tomography, DOT), to account for the fact that the sampling is not along a straight line between the source and detector. Several different approaches have been applied to this so-called 'inverse problem', including using CT or MRI images to segment the image volume into different structural regions and to assign starting values for the optical absorption and scattering to each segment, based on typical values for the tissue class. This then constrains the possible inverse solutions. Most DOT systems have used light sources and detectors placed externally to the body, e.g. around the breast, head or a limb. More recently, DOT devices have been reported for use within body cavities, e.g. within the rectum to image prostate cancer.

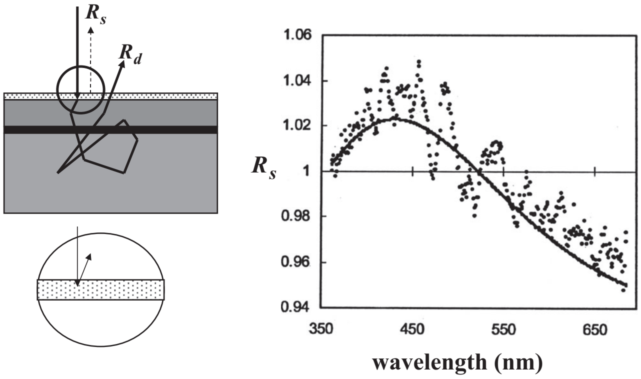
While CW-based transmission imaging is not generally useful, optical tomography using CW measurements may be possible, and there is also much effort in time-of-flight methods or in combining CW and time-dependent measurements. Making measurements at several different wavelengths is also of value, for example, to construct maps of Hb or HbO_2. A further method under development is to use a fluorescent contrast agent that is selected to localize in the tissue of interest (e.g. tumour) and to have a fluorescence lifetime (see below) that is of the same order as the mean photon propagation time. Measuring on and off the fluorescence emission wavelength then provides additional contrast to reconstruct the image.

A recent advance in diffuse imaging is the use of patterned illumination, the so-called spatial light modulation imaging. This is achieved by reflecting the diode laser or LED light beam off a digital micro-mirror array that can then be programmed to project a sinusoidal pattern at different spatial frequencies onto the tissue surface. The demodulation and phase of the diffusely reflected image then allow separation of the tissue absorption and scattering. This has recently been implemented in endoscopic mode. Spatial light modulation also allows sampling of the tissue at different depths, by varying the spatial frequency: the high-frequency pattern mainly samples tissue close to the surface, while lower frequencies are sensitive to depth.

Even under optimal conditions, the spatial resolution achieved to date *in vivo* with diffuse optical imaging more closely resembles that of positron emission tomography, i.e. ~ mm, than it does X-ray computed tomography. However, given the potential for functional/metabolic information that is not available in CT or MR imaging (<100 µm), it is likely that DOT will have specific clinical niches, especially in combination with the other radiological imaging techniques. A case in point is to use breast DOT to monitor whether or not there are metabolic changes in breast cancer during chemotherapy (Figure 22.9b): if there is no marked response, then the drug(s) may not be effective, so that other options may be considered.

22.2.3 Elastic Scattering Spectroscopy

Elastic scattering spectroscopy is a special case of reflectance spectroscopy, as illustrated in Figure 22.9, based on recognition that the information about the mucosal layer in hollow organs (lung, gastrointestinal tract, etc.) is swamped by the large diffuse reflectance from deeper structures. This multiple-scattering background can be markedly reduced either (a) by numerical subtraction using a multi-layer model of the tissues with known or approximate average optical absorption and scattering coefficients for each layer, or (b) by using a polarized light beam and detecting the backscattered light that retains this polarization, thereby blocking the light that comes from the deeper tissues that has lost the polarization through multiple scattering. In either case, this yields the fine structure of the light that has undergone only a few scatterings in the surface layer. By fitting such spectra to a Mie Theory model of light scattering from discrete particles, the distribution of scattering ‘particle’ size can be determined. This is thought to provide information on the size distribution of cell nuclei, which is a very important factor that changes when cells become malignant. This concept has been extended to wide-field imaging, detecting many wavelengths simultaneously or rapidly in succession.



(a) (b)

FIGURE 22.9 Elastic scattering spectroscopy: (a) schematic illustrating the backscatter from nuclei in the top (mucosal) layer of tissue and the background diffuse reflectance from the underlying layers; (b) example of a spectrum from gastro-intestinal tissue, showing the broad diffuse background and the fine structure due to single scattering from the mucosal layer. (After *M. Feld and L. Perelman, USA*.)

22.2.4 Optical Coherence Tomography

Specular reflection (near-180° backscattering) also occurs at micro-scale interfaces within tissues where there is a local refractive index change. This is the basis for optical coherence tomography (OCT), in which sub-surface structural images of high spatial resolution are generated to a depth of a few mm. OCT was developed in the early 1990s, originally for imaging in the eye and is now a standard clinical technique in ophthalmology and optometry, e.g. for retinal imaging that shows the detailed layered tissue structure in a scan time of only a few seconds: Figure 22.10c,d. OCT is the optical analogue of high-frequency ultrasound imaging, in which echo-ranging is used to map the depth distribution of micro-inhomogeneities in the tissue acoustic impedance, from which a 2-D image may be produced by scanning the ultrasound beam laterally. Since the speed of light is so high, direct timing of optical pulses is not practical for spatial resolutions at the required scale of <~10 µm. Hence, an interferometric scheme is employed to provide the spatial localization, using a short coherence length light source. The principle is illustrated in Figure 22.10a. The light is split 50:50 between the tissue and a mirror that scans axially. The light that is reflected specularly (preserving the coherence) from internal optical interfaces then mixes with the reference beam, producing an interference signal that is measured by a photodetector. There is constructive interference, and hence a measurable signal, only when the pathlength in the probe beam and the reference beam are equal (within the source coherence length). Thus, as the reference mirror is scanned, a depth profile of internal reflectance in the tissue is generated. This is analogous to A-scans in ultrasound imaging. The beam is then scanned laterally across the tissue surface to produce the equivalent of a B-scan, as in Figure 22.10d. As shown in Figure 22.10d, the axial and lateral image resolutions in OCT are independent. The former is determined by the coherence length which, in a medium of refractive index n, is given by $\lambda_c^2/(n.\Delta\lambda_c)$, where λ_c is the central wavelength and $\Delta\lambda_c$ is the source bandwidth. Typically, in order to achieve maximum depth of imaging in tissue, NIR sources are used, either in the 800–850 nm range or around 1.3 µm. The latter wavelength is away from the water absorption peak around 970 nm, and there are many components available at this wavelength that have been developed for telecommunications and which can be used for fibre-optic implementation of OCT. For example, using a 1.3 µm source with a bandwidth of about 60 nm, the axial resolution in tissue (assuming an average refractive index of 1.4) is ~10 µm. The lateral resolution is determined primarily by the numerical aperture of the beam, with high NA providing tight focusing (in principle to the diffraction limit) but a limited depth of focus, and *vice versa*: Figure 22.10b. Typically, the lateral resolution is several times worse than the axial resolution in clinical systems.

Since the coherence length is such a critical factor in determining the ultimate resolution in OCT, the availability of light sources with appropriate characteristics is a key issue. The simplest, cheapest and most compact source is a superluminescent diode. Typically, these have a bandwidth of 30–90 nm in the 800–1500 nm range, which may be adequate for some applications, but they have relatively low fibre-coupled output

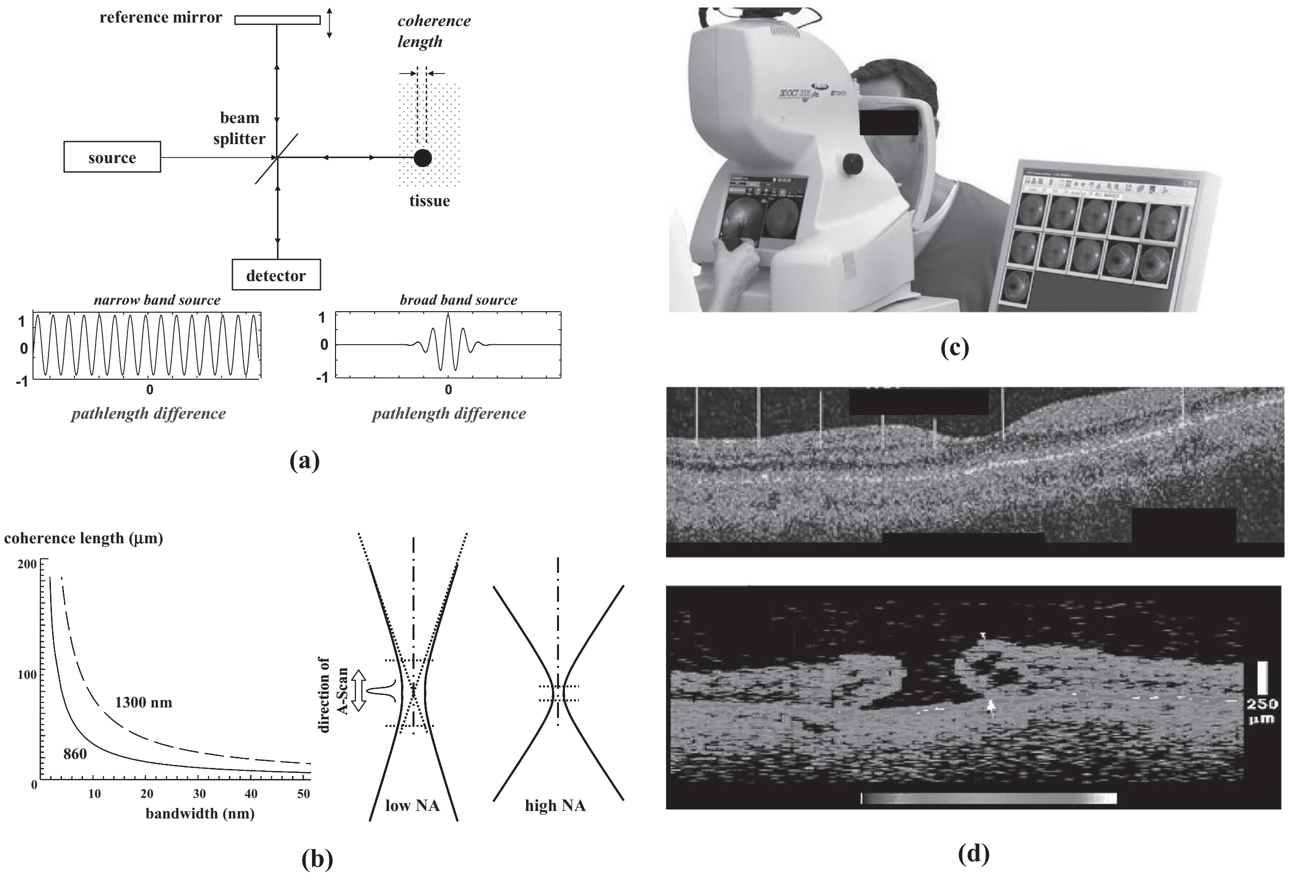


FIGURE 22.10 (a) Principle of OCT, showing the basic Michelson interferometer and the signals obtained from a long coherence-length (monochromatic) source (left) and a short coherence-length (broad band) source (right) as the reference mirror is scanned to vary the pathlength difference between the sample and reference arms of the interferometer. (b) Axial resolution as a function of the light-source bandwidth and the effect of numerical aperture on the lateral resolution and depth of focus. (c) Typical clinical OCT system for retinal imaging. (d) Examples of retinal cross-sectional images showing the various tissue layers: top, normal retina; bottom, retinal hole. (*After NYEEI, USA*.)

power (<5 mW). In order to achieve resolution close to the cellular scale of a few microns, one option is to use a fs pulsed laser, for which the short pulse length gives an equivalent finite bandwidth. Most reports have used a Ti:sapphire laser pumped by a frequency-doubled Nd:YAG laser producing light around 800 nm. For the 1.3 μm range, an alternative is the Cr:Fosterite laser (1280 nm). Such sources have average powers of 10–100 mW and >100 nm bandwidth. Unfortunately, they are not well suited to clinical use, because of size, complexity and cost. The need for a clinically ergonomic source with this type of output is required. It is also the case, however, that this bandwidth exceeds the transmission capabilities of some of the optical elements currently used in fibre-optic OCT systems (see below), and spectral dispersion within the tissue also becomes limiting.

There have been significant technical developments in OCT since the first generation of device time-domain instruments built according to Figures 22.10a and 22.11a, driven in large part by the need for greater imaging speed without sacrificing resolution. Modern devices operate in the Fourier domain and there are two main embodiments, spectral-domain and swept-source. In both approaches the geometry of the reference arm is fixed and depth- ranging is achieved from the spectral properties of the interferogram. The spectral-domain systems use a broad-band (commonly CW) light source and perform spectral analysis through a dispersion grating and (line) spectrometer on the detection end. Hence, the light sources can be quite simple and are commonly based on robust superluminescent diode technology. Conversely, in swept source OCT systems the detection end is quite simple (e.g. a photodiode) and the spectral information is encoded on the source end. Thus, a very narrow spectral line (sub-micron bandwidth) is rapidly modulated across the much wider source bandwidth (typically ~> 100 nm). Sweep speeds determine the A-scan rate and are typically a few hundred kHz for commercial systems. Custom-built research systems are often faster, and (tens of) MHz sweep rate systems have been reported. The swept-source light source designs vary, often amplifying the weak emission from a semiconductor optical amplifier and using various optical methods such as MEMS modulation or polygon ring cavities to shape, narrow and sweep the emission line. Vertical cavity surface emitting lasers offers a slightly different but highly modular and compact solution.

As in ultrasound, imaging of the velocity of moving structures in the tissue is possible. In ultrasound, the Doppler shift can be detected directly, that is, the change in the ultrasound wavelength due to relative motion between the source/detector and the reflecting structure. In the optical domain, the speed of light is so high that the actual colour shift is negligible. Rather, movement of the scattering object along the axial direction results in a phase shift of the interference pattern as the reference mirror scans between successive A-scans. (Hence, the effect is not a true Doppler shift but provides similar information.) This requires that the detailed interference signal is sampled, not just the interference envelop. Since the phase shift can be measured to great precision, the velocity sensitivity can be very high.

In terms of clinical applications, OCT is most advanced in ophthalmology, for high-resolution imaging of the retina, as illustrated in Figures 22.10c and d. For general clinical use beyond ophthalmology fibre-optic implementations of OCT are often required. This involves replacing the classical interferometer by fibre-optic beam splitting/coupling, as seen in Figure 22.11a and development of specific probes for endoscopic or intra-vascular applications as seen in Figure 22.11b. There are many detailed opto-mechanical-electronic designs and operational considerations in OCT instruments and the field continues to develop rapidly in several directions, including for endoscopic and intravascular imaging using either side-looking or circumferentially scanning probes that produce transverse cross-sectional images, using either rotating-wire driven or use MOEMS scanning of the light beam. These probes can be placed axially within a transparent balloon for centreing within hollow structures.

Additional information about the tissue is also accessible, for example, using OCT elastography to image the mechanical properties by applying time-varying pressure to the tissue and monitoring the (micro)structural distortion. This may be particularly useful to distinguish subtle changes associated with, for example, early-stage breast cancer. Deep access within solid organs such as the breast has been enabled by the recent development of 'OCT-in-a-needle' devices, as illustrated in Figure 22.11c.

The signal processing required in current OCT methods is substantial. For example, using spectral OCT within arteries and applying texture-analysis algorithms, it is possible to

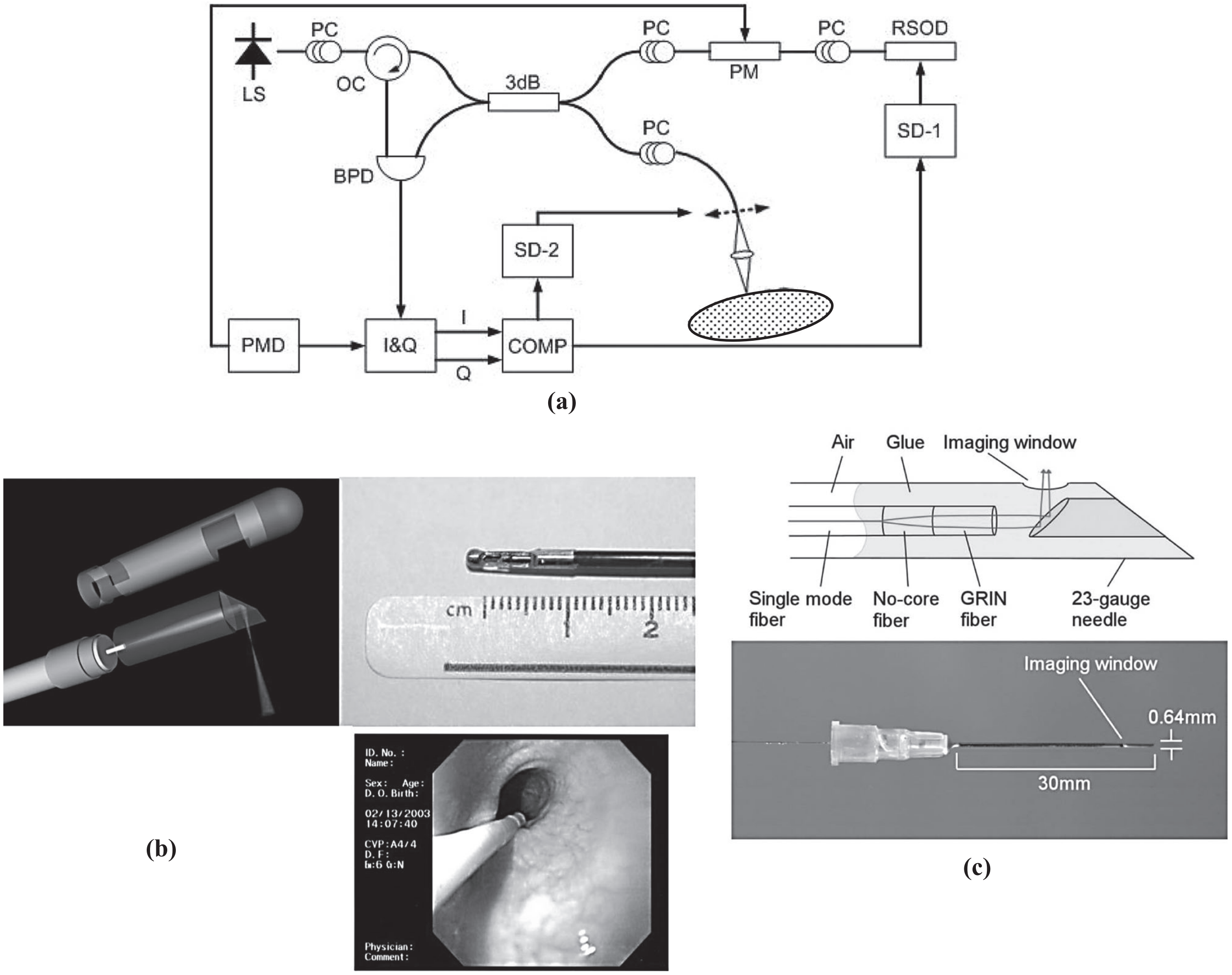


FIGURE 22.11 (a) Example of the schematic of a fibre-optic OCT: LS, light source; PC, polarization controller; OC, optical circulator; PM, phase modulator; RSOD, rapid scanning optical delay; SD, scan drivers; PMD, phase modulator driver; BPD, balanced photodetector; I&Q, in phase and quadrature demodulator; COMP, computer. (b) Example of endoscopic OCT showing a probe with a rotating micro-prism at the end of the optical fibre to direct and collect the light sideways, which is moved axially within a protective sheath. (c) Example of OCT-in-a-needle. (*Courtesy OBEL, Australia.*)

reconstruct maps of the distribution along the artery of important pathological features, such as calcium and fat deposits that are associated with atherosclerosis. Another example is the use of OCT to scan the whole surface of breast lumpectomies, where the clinical objective is to determine if, after surgical resection of the 'lump', there is still tumour at the surface, in which case resection is continued. Tissue-texture classification algorithms are also critical in this application in order to distinguish malignant tumour from normal glandular tissue or abnormal but benign changes. As exemplified by these applications, the volume of data generated in OCT can be extremely large, so that rapid computation is required.

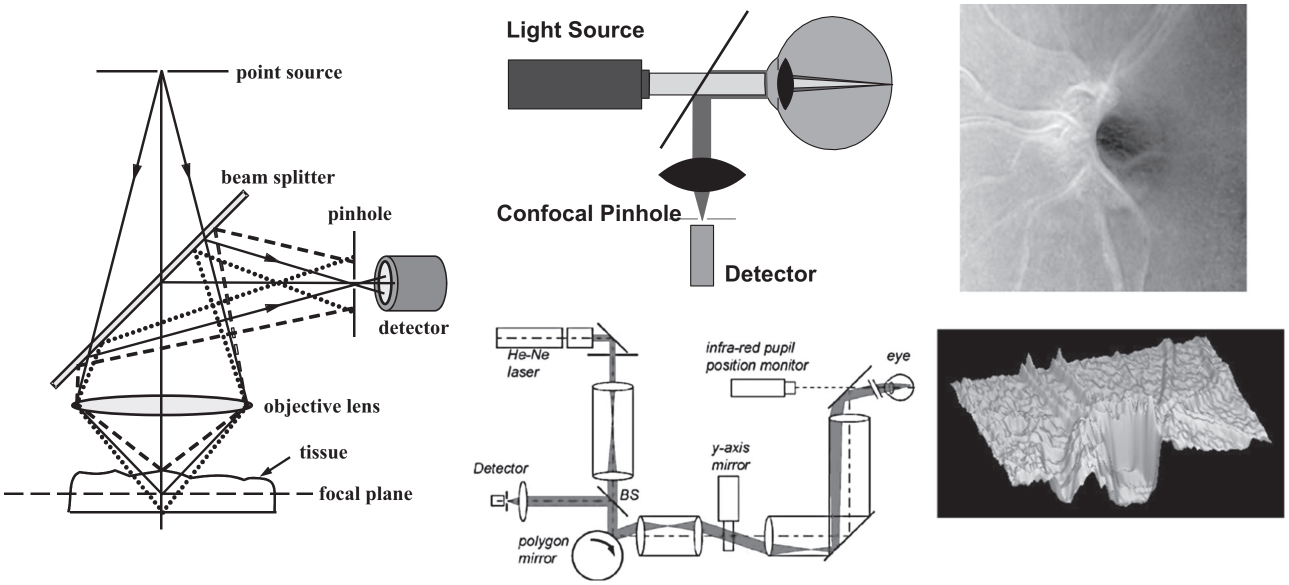
There are many possible future developments in OCT, including further technology transfer from the photonics sector, especially telecommunications; improved short coherence length sources; multiplexing to circumvent the current tradeoff that is needed between resolution, imaging depth and scanning time; the use of contrast agents and multiple wavelengths to enhance image quality and biological information content; specialized probes for interstitial, endoscopic, intravascular and intraoperative use; enhanced Doppler techniques for higher sensitivity and greater dynamic range in blood flow imaging; and signal post-processing and digital image analysis for improved image quality and feature enhancement/identification. A major question (as also with confocal imaging discussed below) is whether or not sufficient depth of imaging and resolution can be achieved to ensure that OCT will be adopted into routine clinical practice for different applications: for example, in endoscopy it is critical to image the boundary between the mucosa and the underlying muscle layers in order to see if there is tumour invasion, since this determines the treatment options. This requires imaging up to a depth of about 3 mm, which is borderline for OCT at this time.

22.2.5 Confocal Imaging

For some applications an alternative to OCT that also provides micron-scale spatial resolution of subsurface tissue structures is confocal optical imaging. Optical 'depth sectioning' is a well-established technique in microscopy, illustrated in Figure 22.12a. The light from a laser is passed through a small pinhole to create essentially a point source. The beam is expanded and passes through a beam splitter, before being focused to a point within the tissue. The reflected light that comes back along the same path is reflected by the beam splitter and focused through a second pinhole onto a sensitive photodetector. Since light coming from the tissue above and below the focal plane is blocked by the detector pinhole, this system defines a (diffraction limited) point within the tissue from which signal is collected. The laser beam is then scanned in the focal plane to produce an X-Y image, while successive (Z) planes are scanned to generate a 3D image.

This concept is being applied also to imaging tissues *in vivo*, including imaging skin tumours to guide surgical resection, in endoscopy ('*in situ* pathology') and, as illustrated in Figure 22.12b, for imaging in the eye (confocal laser-scanning ophthalmoscopy) for which confocality enables both very high resolution (e.g. to image individual photoreceptors) and also depth reconstruction of structures such as the optic nerve head for diagnosis of glaucoma.

Endoscopic implementation of confocal imaging is challenging because of the need to scan the beam at the end of the endoscope. Several alternative approaches are possible: direct miniaturization of confocal microscope system; the use of MOEMS devices, for example, scanning micro-mirrors at the end of the endoscope; rapid mechanical rotation of the OCT probe and; using a fibre-optic endoscope, exploiting the



(a) (b)

FIGURE 22.12 (a) Principle of confocal imaging, showing the detection of light from the focal plane (solid lines) blocking of light rays from above and below the focal plane (broken lines). (b) Confocal laser-scanning ophthalmoscopy, showing the principle, a schematic of a high-resolution, real-time instrument that uses a rotating polygonal mirror for fast scanning in one plane that enables video-rate imaging, and 3D images of the optical nerve head obtained by reconstructing a stack of confocal images at different depths. (*Courtesy M. Campbell, Canada.*)

fact that each fibre acts as a detection pinhole and so collects light at a specific sub-surface depth, the image being created by scanning the input laser beam across the other end of the fibre bundle.

Compared with OCT, confocal imaging *in vivo* can provide higher resolution (<~μm) but generally has less depth of imaging (<~500 μm). In addition to operating in reflectance mode, confocal imaging can be based on detection of either the fluorescence or Raman signals from tissue (see below), which allows the possibility of obtaining functional/molecular information in addition to the tissue ultrastructure.

As is already established in microscopy, future developments are likely to include the use of 2-photon confocal imaging *in vivo*, using a fs laser source to generate multi-photon interactions at the focal point within the tissue. This improves the resolution (through the quadratic dependence on intensity) and reduces photobleaching and possible damage to the tissue through the use of NIR rather than visible illumination. It may also allow an increase in the depth of imaging. Cost of the hardware remains an issue with confocal technologies, and this is an aspect where new laser sources and micro-optical devices will be critical.

22.2.6 Fluorescence Spectroscopy and Imaging

Fluorescence has been investigated for a wide range of potential clinical applications. It is, of course, used frequently as a research tool in biology, as in fluorescence microscopy, flow cytometry and many bioassays. In fluorescence, short wavelength (typically ultraviolet or blue) light is absorbed by specific molecules (fluorophores) in the tissue to generate electronic excited states that can de-excite through emission of a spectrum of longer wavelength light. Although ~10^3 times weaker than the diffuse elastic-scattering signal from tissue, the fluorescence light is easily measured using high-efficiency optical filters to reduce the elastically scattered background light and high-sensitivity photodetectors or photodetector arrays for point spectroscopy or fluorescence imaging, respectively: a point spectroscopy system schematic is shown in Figure 22.13a.

The importance of fluorescence-based techniques *in vivo* lies in the sensitivity to biochemical (and physiological and structural) tissue features, particularly within the lining of hollow organs such as the lung, gastrointestinal tract, bladder, oral cavity and cervix, which are sites for early cancer and pre-malignant changes in tissue. These can be detected either by the changes in autofluorescence, as illustrated in Figure 22.13b, or by preferential uptake of an exogenous fluorophore (that has been administered to the patient), for example, for imaging the retinal blood vessels following intravenous injection of fluorescein dye or for imaging tumours using a wide variety of agents. Fibre-optic-based fluorescence spectroscopy has been explored in the past for intra-vascular guidance in laser angioplasty to clear blocked arteries.

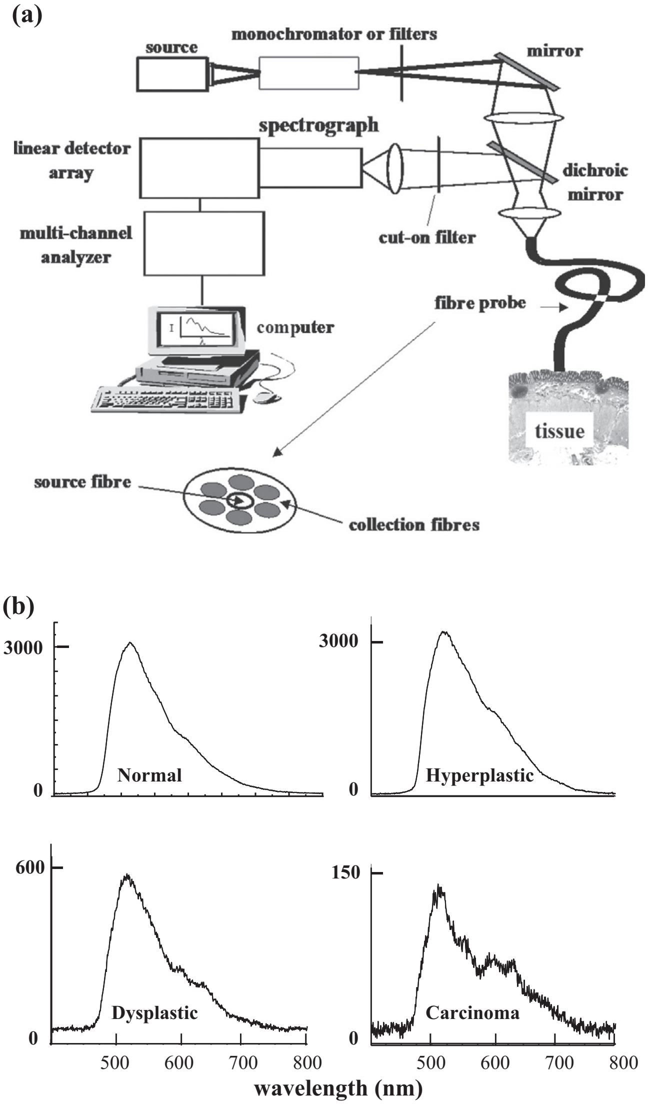


FIGURE 22.13 (a) Schematic of fibre-optic fluorescence spectrometer. (b) Example of spectra measured at points within the colon during endoscopy (excitation wavelength 430 nm), where the fibre probe is placed through the instrument channel of the endoscope in gentle contact with the tissue: different stages of cancer development are seen, illustrating the diagnostic power of this technique.

An increasingly important application of fluorescence imaging/spectroscopy is to guide surgery. One example is the imaging of intravenously injected of the fluorophore indocyanine green, which stays in the vasculature, to determine if tissues have adequate blood perfusion for transplantation. Other fluorescent agents that have specific disease targeting, particularly for cancer, are also being used to determine tumour margins during surgery. This is most mature in resecting malignant brain tumours (gliomas), by incorporating fluorescence capability into a neurosurgical microscope. A recent extension of this technique is to actually quantify the (absolute) concentration of the fluorophore in tissue by parallel use of diffuse reflectance imaging/spectroscopy to determine the tissue absorption and scattering coefficients at the fluorescence excitation and emission wavelengths. These are then used to correct the measured fluorescence signal for the light attenuation. Using longer wavelength excitation together with spatial-light-modulation in both diffuse-reflectance and fluorescence modes allows fluorophore quantification and ‘topographic’ profiling of sub-surface tumour, up to depths of 10 mm or greater if near-infrared fluorophores are used.

In the past few years, various 'point-of-care' fluorescent imaging systems have been developed: one example is shown in Figure 22.14 comprising a hand-held device based around a cell-phone camera, with additional LED light sources and optical filters. This is sensitive enough to detect the autofluorescence from bacteria on or near the tissue surface and is now being used to guide 'clean up' of chronic infected wounds. It may also be applied to other surgical-guidance applications, including in cancer. This figure also illustrates the difference between a prototype optical diagnostic device and a commercial 'product' for routine use in patients.

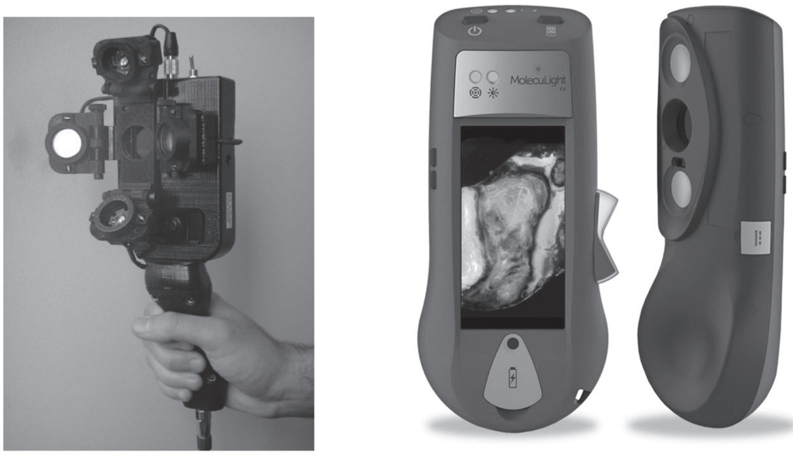
With exogenous fluorophores the optimal excitation and emission wavelengths are known and the signal is usually much higher than with autofluorescence, so that lower excitation power and/or less sensitive detectors can be used. Although it is possible, in principle, to use a single excitation and detection wavelength to detect an exogenous fluorophore, it is difficult to make this reliable and quantitative. For example, the autofluorescence background also contributes to the signal and can vary from point to point or from patient to patient, as can the tissue attenuation. Hence, it is more common to measure at several wavelengths around the peak and/or to use more than one excitation wavelength to separate the fluorophore signal from the autofluorescence background.

Although point fluorescence spectroscopy can provide valuable diagnostic information, there has also been considerable effort to develop fluorescence also as an imaging technique. For accessible external surfaces, such as the skin, this can be straightforward. The excitation light is spread as a uniform beam over the area of interest and, using optical filters, fluorescence images captured at one or more wavelengths of interest that may then be combined. Endoscopic imaging is technologically much more challenging, especially since it usually has to be integrated into existing white-light instrumentation. First, the excitation light must be coupled into the endoscope. Second, one or more additional detectors must be used since, due to the relatively low fluorescence intensity and the poor optical efficiency of fibre-optic endoscopes, the CCD cameras used for the white-light imaging are usually not sensitive enough for real-time fluorescence imaging. Third, the cost and complexity increases if more than one detection wavelength is used, since this requires either multiple (intensified) cameras or rapid wavelength switching with a single camera. Nevertheless, practical clinical systems have been developed. For example, one commercial fluorescence endoscopic system uses an intensified-CCD camera to detect the green fluorescence from the tissue under blue light excitation, simultaneously with a red reflectance image. The excitation light is obtained from the white-light source by rapid insertion of a filter, at the same time as a shutter is used to switch to the ICCD and a red-pass filter in the detector head. The system provides real-time (video rate) fluorescence images, in which the red reflectance and green fluorescence channels are combined into a false-colour image.

In terms of the further development of fluorescence diagnostics, likely trends include broadening of the applications, for example, to guide surgery and other therapeutic modalities as well as monitoring treatment responses. Fluorescence-lifetime imaging is also an emerging modality that has the advantage that the lifetime is independent of the fluorophore concentration and provides information on the tissue microenvironment which affects the binding of the fluorophore and hence its decay rate following a short light pulse, e.g. from a ps gain-switched diode laser operating at tens of MHz. There is also increasing use of novel fluorescence dyes, including near-infrared compounds in order to reduce the tissue autofluorescence background signal; fluorescent nanoparticles; fluorophores linked to disease-specific targeting agents such as antibodies; and 'molecular beacons' that become fluorescent (unquenched) only upon binding to or interacting with a specific biological substrate. All these approaches are intended to improve diagnostic sensitivity and specificity.

22.2.7 Raman Spectroscopy and Imaging

In Raman, i.e. inelastic, scattering a photon interacting with a molecule can either lose energy and be shifted to a longer wavelength (Stokes scattering) or gain energy (anti-Stokes),



(a) (b)

FIGURE 22.14 Cell phone camera-based fluorescence device used for imaging of bacterial autofluorescence in infected wounds. (a) Prototype used for initial clinical studies showing the LED illumination sources. (b) Subsequent commercial device. (*Courtesy Moleculight, Canada.*)

resulting in a corresponding change in the vibrational or rotational state of the molecule: Figure 22.15a. The former is normally used in diagnostic applications, since the signal intensity is significantly higher than in anti-Stokes scattering. Raman spectra typically show multiple narrow peaks (~1–3 nm) rather than the broad structures seen in fluorescence. The spectra are characteristic of stretching and bending modes of common molecular bonds such as lipids, proteins and nucleic acids, and so provide a characteristic 'molecular fingerprint'. However, specific molecular species are not usually identifiable.

The availability of compact NIR diode lasers of adequate power (~100–500 mW), high-throughput holographic spectrographs and high-sensitivity cooled CCD array detectors, together with special-purpose optical fibre probes, has enabled *in vivo* Raman instruments, as shown in Figure 22.15c. These systems typically have spectral resolution of ~4 cm^{-1}. The key to this technology is the use of small optical filters built onto the distal end of the delivery and collection fibres. On the delivery fibre, a cut-off filter just above the laser excitation wavelength is used to prevent fluorescence or Raman light that has been generated in the fibre itself from being launched into the tissue. On the tip of the detector fibres, a cut-on filter is used to block the elastically scattered light from the tissue from entering the fibre and subsequently generating more fibre fluorescence and Raman light. This leaves mainly the tissue autofluorescence and Raman signals. While the former is 2–3 orders-of-magnitude stronger, its spectrum is broad and relatively unstructured, so that it can be substantially subtracted after applying a low-order polynomial fit to the measured spectrum: Figure 22.15b. Additional filtering at the detector end of the fibres further reduces the background, leaving a clean Raman spectrum. Varieties of these instruments have been tested in various clinical applications, including during endoscopic imaging for cancer detection and localization and for 'optical biopsy', either to replace or complement conventional techniques. The last has recently been demonstrated to guide surgery or biopsy of brain and other tumours.

Changes in the tissue with (early) disease do not generally result in obvious Raman spectral changes, such as loss or gain of specific peaks. Rather, as seen in Figure 22.15d, the changes are subtle and distributed across the spectrum, so that simple ratiometric algorithms that work well in fluorescence mode are usually inadequate. The so-called chemometric techniques, such as Principal Component Analysis or neural networks, are required, in which a set of spectra of known tissue pathology are used to 'train' the algorithm, after which an unknown spectrum is presented as input and the algorithm makes a predictive classification decision (e.g. normal or abnormal). There are many subtleties in these procedures to guard against bias but they can provide highly reliable results. The training and testing process has to be established for every clinical condition of interest.

Raman spectroscopy *in vivo* has also been implemented in confocal geometry to provide sub-surface information, although the scanning times are long because of the weak signals. Future developments that would be valuable

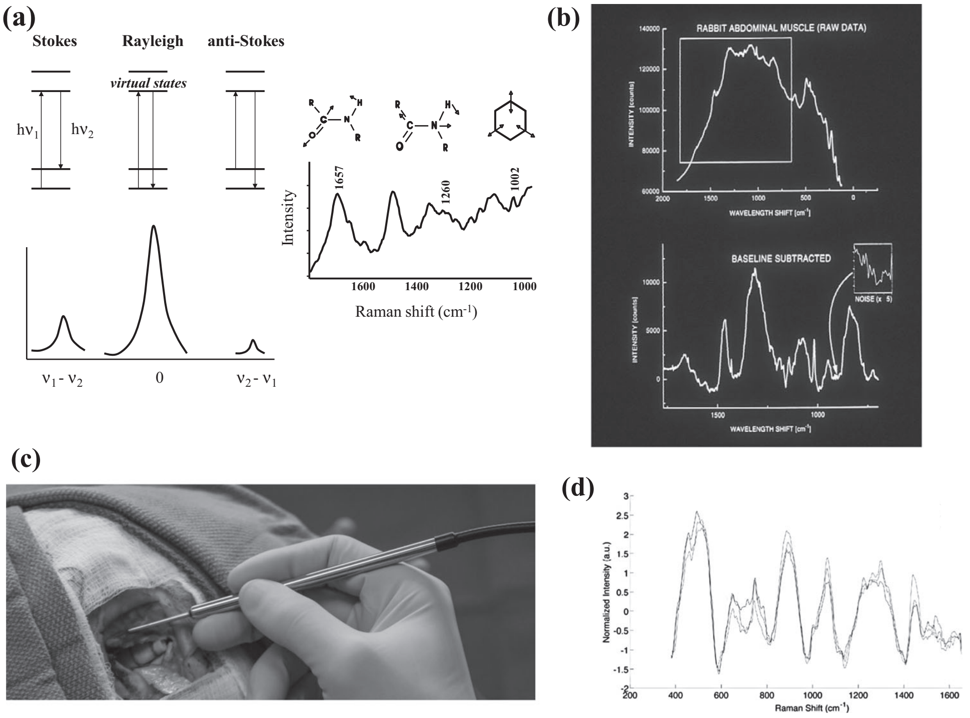


FIGURE 22.15 Raman spectroscopy. (a) The principle of Raman scattering: note that the relative heights of the Raman and elastic (Rayleigh) peaks are not to scale, the latter being much larger than shown. A typical NIR Raman spectrum of soft tissue, shown on the right, with 3 vibrational molecular bonds identified. (b) Spectra measured with fibre-optic Raman instrument: top, -before; bottom, after subtracting the fluorescence background: only the spectral range above about 500 cm^{-1} is used for diagnosis since the shorter wavenumbers are corrupted by Raman background from the fibre that cannot be easily removed. (c) Fibre-optic probe of a commercial Raman system used for guiding tumour surgery and representative spectra from the normal brain and brain tumour showing the subtle spectral differences with disease. (*Courtesy ODS Medical, Canada.*)

include cheaper and more compact instrumentation (source-spectrometer-detector) and specialized Raman probes with low spectral contamination, either through the use of new fibre materials and/or incorporation of efficient spectral filtering. There are also several ways to increase the Raman signal to enable wide-field Raman imaging *in vivo*. The first is surface-enhanced Raman spectroscopy (SERS), exploiting plasmonic amplification of the molecular signal at the surface of metal nanoparticles (gold or silver). The nanoparticles are either in direct contact with the tissue or Raman-active molecules are linked to their surface. An advantage of the latter approach is that it allows multiplexed detection, using a 'cocktail' of nanoparticles with different combinations of reporters and targeting antibodies, as illustrated in Figure 22.16 The resulting spectrum in each image pixel is then the sum of all the Raman spectra from the different reporters, weighted by their relative concentration, and can then be spectrally unmixed to give the concentrations at each point in the image. Up to 10-plex imaging has been demonstrated in this way. Such bio-multiplexing can aid in personalizing treatment, particularly in cancer.

The second approach to Raman bio-imaging is to use either coherent anti-Stokes scattering (CARS) or stimulated Raman scattering (SRS). These are 2-step processes that require short-pulse laser sources. CARS differentially amplifies the strongest signals such as from lipids, while in SRS, the spectra amplified uniformly to give undistorted spectra that have same information content as spontaneous Raman, only several orders of magnitude stronger. Both techniques require pulsed lasers with particular specifications. SRS can be done with either fs or ps pulses. For example, SRS imaging has recently been reported for guiding tumour surgery using the so-called Stokes beam at 1064 nm (7 ps, 76 MHz) and a tunable pump beam from an optical parametric oscillator, temporally and spatially overlapping the two pulse trains. SRS is increasingly likely to be used instead of fluorescence if the cost and complexity of the laser sources can be reduced. A further advantage of SRS imaging is that small molecular-weight reporters can be used that have Raman peak(s) that are well separated from the endogenous Raman signals. Because of their small size, these molecules do not interfere with the cellular functions, which is often not the case with much larger fluorescent reporters.

22.2.7.1 Photoacoustic Imaging

Photoacoustic (or optoacoustic) imaging (PAI) combines the molecular specificity of optical interactions with the depth capability of acoustics. As illustrated in Figure 22.17a, the principle is that, upon absorption of pulsed laser light at a

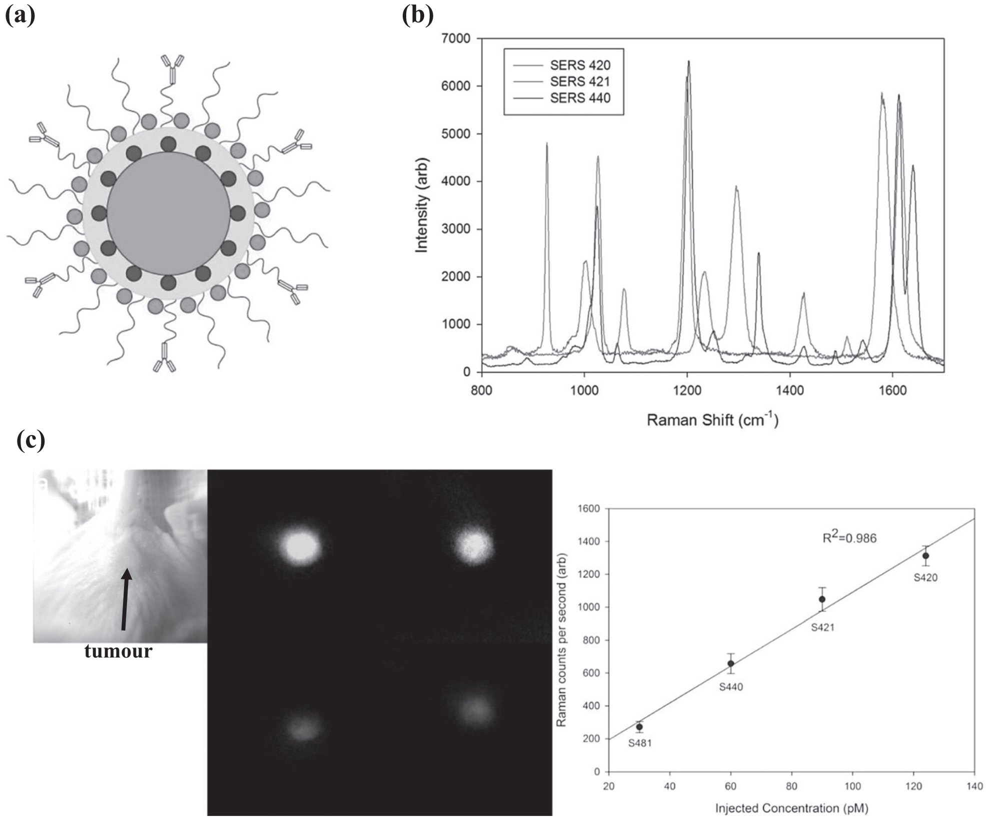


FIGURE 22.16 Example of multiplexed SERS imaging. (a) Gold nanoparticle coated with specific Raman-active reporter molecules and with antibodies attached to the surface for tumour targeting. (b) Overlapping Raman spectra from a 'cocktail' of 4 different SERS nanoparticles each with a different reporter molecule and antibody. (c) Imaging in a tumour model injected with the 4-plex cocktail. The 4 'spots' each correspond to a specific SERS reporter after unmixing the spectra measured in each image pixel (the images have been spatially separated for clarity from a single composite image), while the graph of the SERS signal *versus* the injected nanoparticle concentration shows that the unmixing is quantitative, with each component correctly represented in the images.

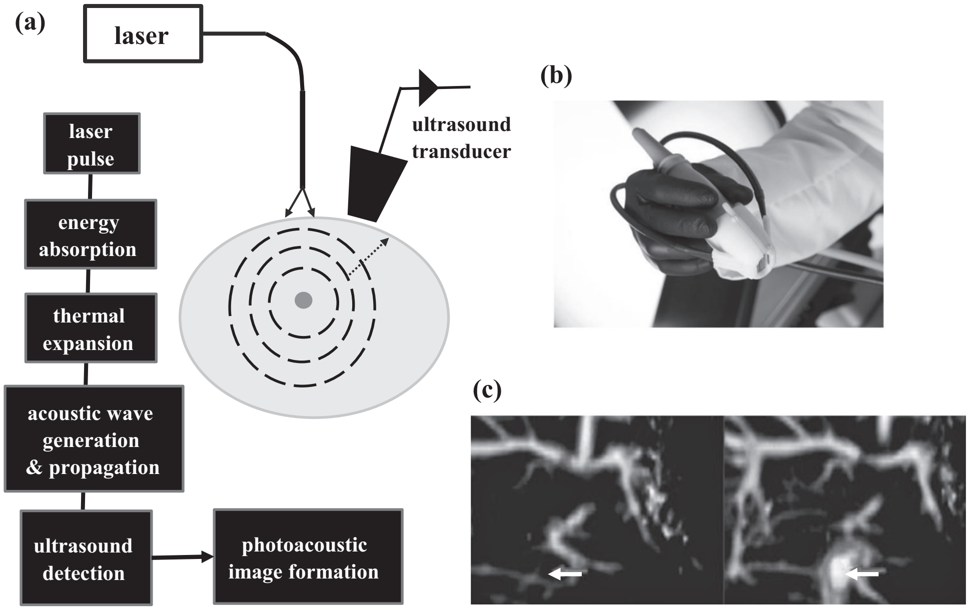


FIGURE 22.17 Photoacoustic imaging. (a) Principle. (b) Example of integrated ultrasound transducer and optical fibres for pulsed laser delivery. (c) Example of photoacoustic imaging *in vivo* before (left) and after (right) injection of nanoparticles into the lymphatic system: the arrow points to a lymph node that is highlighted by the nanoparticle absorption, while the other structures are blood vessels highlighted by the absorption of the laser pulse by haemoglobin.

point within the tissue, there is rapid localized transient heating that causes thermal expansion and generates an acoustic wave that travels to the tissue surface where it is detected using an ultrasound transducer. The acoustic pressure generated at depth in the tissue is proportional to the product of the local light fluence, optical absorption coefficient and the so-called Grüneisen parameter, which is a tissue-dependent measure of the photoacoustic conversion efficiency. Hence the measured acoustic signal is given by this pressure and by the attenuation of the acoustic wave as it propagates to the detector, which is usually on the tissue surface. The spatial resolution and detection depth in tissue that, respectively, increase and decrease with increasing frequency, are then those of the acoustic wave. Imaging up to several cm through tissue has been reported. (Photoacoustic microscopy achieves much higher spatial resolution but only in thin tissue specimens.) The acoustic signals are typically broadband, which impacts the ultrasound transducer design. The light source for PAI is typically a Q-switched, pulsed Nd:YAG laser providing several tens or hundreds of mJ and pumping broad-band optical parametric oscillators, giving tunable vis/NIR ns pulses. There is considerable interest in less expensive and more compact laser sources for PAI, including diode lasers, the challenge being to achieve high-enough peak pulse power in the optimum pulse-length and pulse-repetition-frequency ranges. Various geometries for the light delivery and acoustic detection have been implemented, depending on the application: an example is shown in Figure 22.17b. These include simple planar geometry for accessible tissue surfaces, arrays of multiple fibres, transducer arrays to image 3D tissue specimens and ‘probes’ for imaging within body cavities.

Unlike other optical diagnostic techniques, the high light scattering in tissue does not affect the spatial resolution, but rather serves to spatially distribute the light throughout the imaging volume. Many studies have been reported where the vasculature in tissue is imaged through the high absorption by haemoglobin (Hb or HbO_2), depending on the wavelength used. Absorbing contrast agents may also be used to selectively enhance the signals from specific biological structures. An example of PAI of using organic porphyrin-lipid nanoparticles is shown in Figure 22.17c. The melanin granules in melanomas also provide marked absorption contrast.

22.2.9 Other Methods

There are several other optical techniques in place or under development for clinical diagnostics, including the following.

a. speckle interferometry, in which the time-varying speckle pattern of laser light reflected from tissue is used to provide information on, for example, blood flow.
b. fluorescence correlation spectroscopy, in which the dynamic variations in the local concentration of fluorophores within a small probed volume of tissue may give information on, e.g., diffusion constants.
c. photothermal spectroscopy, in which the time-dependent change in (black body) infrared emission from the tissue surface following energy absorption (and hence temperature rise) due to an incident light pulse is used to provide depth profiling, for example,

to guide laser treatment of vascular malformations in skin (‘birth marks’).

d. CW laser Doppler blood flow measurements, based on detecting the small shift in wavelength of a monochromatic (laser) source.
e. polarimetry to detect chirally-active molecules such as glucose in the tissue and to map changes in (collagen) structure that can be used to identify tumour margins either alone or in combination with mass spectroscopy, or the use of cross-polarization to enhance sub-surface features in skin by reducing the background multiple scattering.
f. Terahertz imaging using wavelengths in the mm range that may provide different biological information than in the visible or near-infrared range: a limitation for *in vivo* imaging is, however, the high signal attenuation and poor tissue penetration in this wavelength range due to high water absorption.
g. Non-linear techniques that are being translated from established and emerging microscopy applications into *in vivo* clinical instruments.

In addition, several of the spectroscopic and imaging techniques that are currently used in or being developed for studies in cells and animal models may in the future find clinical application. One field that has not been discussed here is the use of X-ray lasers. This is an evolving technology, so that its role in medical diagnostics is not yet clear. However, the possibility to have a radiation source with the tissue penetration of X-rays and some of the properties of lasers, such as coherence and monochromaticity, could open exciting new diagnostic imaging techniques.

22.2.10 Optical Imaging in Pathology

While the above techniques are primarily for *in vivo* use, advances in technologies for histopathological imaging and analysis are also highly relevant. Even although it is often considered as the gold standard against which other techniques are compared, pathologic diagnosis is far from perfect and there may be considerable uncertainty about, for example, the presence and staging of cancer. Hence, novel complementary optical imaging modalities are of increasing value. These include digital microscopy, including rapid whole-slide scanning to facilitate sharing and archiving of images, as well as powerful image processing and analysis techniques to ‘extend the pathologist’s vision’; multi-wavelength band imaging to increase image contrast and help distinguish features of interest; and a variety of non-linear techniques that allow label-free imaging of unstained tissue samples. The last includes second harmonic generation imaging of tissue collagen, with or without polarization sensitivity that provides information at the ultra-structural level, 2-photon fluorescence microscopy and SRS microscopy.

In addition, the so-called laser capture micro-dissection allows specific areas of tissue from a histology slide to be mapped and separated from the rest of the tissue. This enriches the sample in the cells of interest (e.g. cancer) so that further genomic or proteomic analysis can be carried out on them with reduced background from non-target cells. There are several ways to carry out the separation, including using a CW laser beam to locally heat an overlying film that then melts onto the tissue that can then be pulled off or using a pulsed laser to mechanically eject (‘catapult’) the selected cells.

In addition to microscopy, flow cytometry, in which the scattering and fluorescence from single cells flowing within a fluid stream are analyzed, continues to be an important method for sophisticated pathological diagnosis. Other forms of cytology utilizing optical imaging are also important, the best-known example of which is the PAP test used to screen for early cervical cancer by analyzing various morphological and biochemical features of cell samples. It is likely that cell phone camera-based devices will start to be applied to the various cytology techniques, because of their low cost and wide availability, especially in remote or resource-limited locations. Hyperspectral techniques, in which images are taken simultaneously at many different wavelengths, are also emerging strongly because of the additional biological information that can be extracted. New camera technologies, such as mosaic CCD arrays, could play key roles here, especially if the cost can be brought down.

22.3 Comparison of Techniques and General Future Trends

As stated earlier, not all of the above approaches will become routine clinical tools, while new methods will undoubtedly appear. A few general points are pertinent to potential future developments.

1. New concepts and devices from other areas of photonics will continue to enable the biomedical applications, just as many of the components developed for optical telecommunications were directly applicable to fibre-optic-based OCT.
2. Major trends in life sciences and medicine will open new opportunities for photonic solutions. These include ‘personalized’ medicine at one socioeconomic extreme, requiring sophisticated diagnostic systems, and low-cost point-of care technologies at the opposite end. Optical devices are particularly suited to the latter because of their relatively low cost, size and safety.
3. The development of new optical tracers for disease-specific targeting will have major impact, both on the applications and in the optical hardware and analysis required. This includes the different classes of optically-active nanoparticles and other ‘smart’ optical reporters that enable multiplexed and multimodal detection.
4. Conversely, there is increasing interest in multimodal ‘label-free’ imaging that does not utilize exogenous tracers, including non-linear techniques.

5. Algorithms for spectral and image analysis will have to become much more sophisticated and powerful than has been the case to date in optical diagnostics, in order to exploit fully the information content of optical measurements. Machine learning and artificial intelligence will increasingly be needed in optical diagnostics to fully exploit the rich bio-information that is generated.
6. Clinical diagnostics has many different faces: detection of disease or agents of disease, staging, guidance of therapies, follow-up to monitor treatment outcome, and even to prevent disease or detect toxins and pathogens in the environment and *in vivo*. Each domain will require different technological solutions.
7. There will be competing non-optical technologies and, in some cases, several different potential optical technologies, for particular applications.
8. Conversely, for many applications, it may be necessary to use more than one optical technique or to combine an optical technique with a non-optical method to achieve the required clinical impact, exploiting complementary structural, physiological or molecular/genetic information.
9. Theranostics, the term used to describe the integration of diagnostics with treatment is rapidly emerging.

Finally, there is increasing debate about the challenges of translating biomedical optics from the concept stage, through prototyping, pre-clinical testing and clinical trials into commercialization and routine clinical use. This raises many issues: scientific, technological, clinical, ethical, cultural and socioeconomic. Nevertheless, it should be clear that optical diagnostics is a rich field for multi-disciplinary research and development that includes physics, (bio) engineering, chemistry and material sciences, mathematics and computation, as well as the ‘end user’ life scientists and clinicians. It is indeed an exciting and challenging field.

BIBLIOGRAPHY

The following sources serve as useful starting points into the literature on optical diagnostics.

Boudoux C., *Fundamentals of Biomedical Optics*. Pollux Editions, Blurb, Montreal, 2016.

A general graduate- or advanced undergraduate-level textbook on biomedical optics, including spectroscopy and imaging.

Drexler W. and Fujimoto J.G., *Optical Coherence Tomography: Technology and Applications*. Springer, New York, 2015.

A general text on the principles, devices and biomedical applications of OCT.

Georgakoudi I., Jacobsen B.C., Van Dam J. et al., Fluorescence, reflectance, and light-scattering spectroscopy for evaluating dysplasia in patients with Barrett’s esophagus. *Gastroenterology* 120, 1620–1629, 2000.

Useful for considering why and how to combine optical diagnostic techniques.

Lewis I.R. and Edwards H.G.M. (eds.). *Handbook of Raman Spectroscopy*. Marcel Dekker Inc., New York, 2001.

A comprehensive text on Raman spectroscopy, including chapters on biomedical applications.

Mahadevan-Jansen A. and Richards-Kortum R., Raman spectroscopy for the detection of cancers and precancers. *Journal of Biomedical Optics* 1, 31–70, 1996.

A detailed review of Raman principles and techniques, particularly for cancer detection.

Murphy D.B. and Davidson M., *Fundamentals of Light Microscopy and Electronic Imaging*. Wiley-Blackwell, Hoboken, New Jersey, 2013.

Basic principles and instrumentation and probes for optical microscopy.

Mycek M.-A. and Pogue B.W. (eds.). *Fluorescence in Biomedicine*. Marcel Dekker, New York, 2003.

A general text on fundamentals of biomedical fluorescence and its applications.

Tuchin V.V. (ed.). *Handbook of Optical Biomedical Diagnosis*. SPIE, Bellingham, Washington, D.C, 2002.

A major text on the principles and practice of optical diagnostics, with in-depth chapters on many of the topics discussed above and extensive reference lists in each chapter.

Vo-Dihn T. (ed.). *Biomedical Photonics Handbook*. CRC Press, Boca Raton, FL, 2003.

A multi-authored text that includes several detailed chapters on medical diagnostic techniques and instrumentation.

Wang L.V., *Photoacoustic Imaging and Spectroscopy*. CRC Press, Boca Raton, FL, 2017.

A detailed text on photoacoustic techniques and applications.

23

Broad Bandwidth Light Sources in Optical Coherence Tomography (OCT)

Jinxin Huang and Jannick P. Rolland

CONTENTS

23.1 Introduction to OCT 309
    23.1.1 Time Domain OCT 310
    23.1.2 Fourier Domain OCT 310
23.2 Broadband Sources for OCT 312
    23.2.1 Femtosecond Lasers 312
    23.2.2 Swept Source Lasers 313
    23.2.3 Supercontinuum Sources 313
    23.2.4 Superluminescent Diodes 314
23.3 Source Parameters and Their Indications for OCT Performance 314
    23.3.1 Spectrum (Central Wavelength and Bandwidth) 314
    23.3.2 Source Noise 315
    23.3.3 Instantaneous Linewidth of a Swept Source 316
    23.3.4 Repetition/Sweep Rate of Broadband Sources 317
23.4 Conclusions 317
References 317

23.1 Introduction to OCT

Optical coherence tomography (OCT) is an optical imaging modality that utilizes the principle of low-coherence interferometry (Huang et al. 1991). Since the first successful demonstration of cross-sectional imaging in the early 1990s, OCT technique has advanced rapidly in the past three decades and has become a unique imaging modality that provides three-dimensional volumetric images with resolutions on the order of microns. OCT can function as a non-invasive biopsy tool and can provide images of tissue both *in situ* and *in vivo* with real-time visualizations. Due to the non-invasive nature and unprecedented axial sectioning capability, OCT has found a wide range of applications in different biomedical fields, including ophthalmology, dermatology, dentistry, cardiology, urology, neurology, laryngology, and gastrointestinal diseases (Drexler and Fujimoto 2008).

To understand where OCT stands in the field of biomedical imaging, Figure 23.1 compares the imaging depth and the resolution across different imaging modalities. On one end of the spectrum, there are imaging techniques with high penetration depth and relatively low resolution, such as X-ray computed tomography, magnetic resonance imaging, and ultrasound imaging. The spatial resolution of these techniques is limited from tens of micrometres to a few millimetres. On the other end, there are imaging techniques with high spatial resolution (<1 μm laterally), such as confocal microscopy, fluorescence microscopy, and two- or multi-photon microscopy. However, the imaging depth of those techniques is typically limited to be less than a few hundred micrometres. With a lateral resolution of about a few micrometres and an imaging depth of a few millimetres, OCT fills a niche among different biomedical imaging modalities.

OCT imaging operates with the physics principle of low-coherence interferometry, which achieves the axial sectioning capability through an interferometer structure with a broad bandwidth light source. In the implementation of an OCT system that performs a point-to-point imaging, the broadband light emitted from the source focuses and illuminates on the sample surface; the backscattered or back-reflected light from the internal microstructures in materials and biological systems is collected into the interferometer. The time-of-flight of the collected light, which encodes the depth profile of the sample under test, is measured through interference. The acquisition of the depth profile at one lateral point is called one A-scan, which represents one depth scan along the axial direction of the beam. When the light beam scans laterally across the sample or if the sample is translated with stages, a 3D volumetric image is acquired. OCT has the main advantage compared to other microscopy techniques to provide unprecedented axial resolution in the micrometre class set by the low temporal-coherence of the broadband source; the development of broadband sources plays a key role in the advancement and wide adoption of OCT.

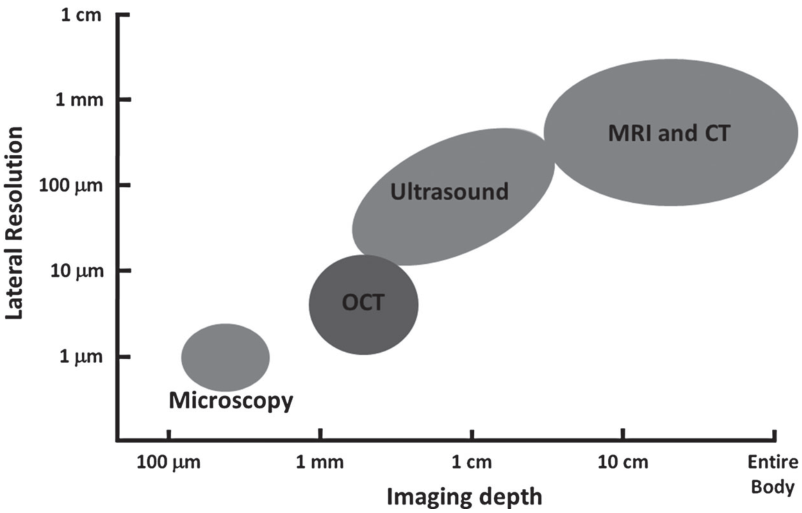


FIGURE 23.1 The imaging depth and lateral resolution of different biomedical imaging modalities. (Adapted from (Drexler and Fujimoto 2008).)

Depending on how the depth scan is performed, there are two types of configurations of OCT systems, time domain OCT (TD-OCT) and Fourier domain OCT (FD-OCT); the time domain configuration performs the depth scan through the mechanical scan of a reference mirror, while the Fourier domain configuration obtains the depth profile through the acquisition of the interference across wavelengths. Sessions 23.1.1 and 23.1.2 will discuss the different configurations of OCT systems.

23.1.1 Time Domain OCT

The early OCT systems were based on the time domain configuration, in which a photodiode is adopted as the detector and a moving mirror is placed in the reference arm, as schematically shown in Figure 23.2a. Due to the temporal coherence gating of low-coherence interferometry, the interference signal is obtained only when the optical path length (OPL) of the back-reflected light from the sample arm and the reference arm are matched. By moving the reference mirror axially, different interfaces of the sample along the depth can be observed, which is then used to reconstruct the depth profile for one A-scan.

To illustrate the TD-OCT imaging process, Figure 23.2b shows one example of the reflectivity profile of a sample; as the reference mirror scans axially, the low temporal-coherence property of the light source behaves like a gating and interference is observed when the OPL difference is within the coherence length of the light source. One example of the recorded signal is illustrated in Figure 23.2c, and the depth profile of the sample is reconstructed as shown in Figure 23.2d. The strength of different layer interfaces, i.e. the changes in refractive indices, is represented by the amplitude of the reconstructed signal in Figure 23.2d.

23.1.2 Fourier Domain OCT

FD-OCT was first described by Fercher *et al.* in 1995 (Fercher et al. 1995). In a FD-OCT system, the coherence gating involves acquiring the interferometric signal as a function of

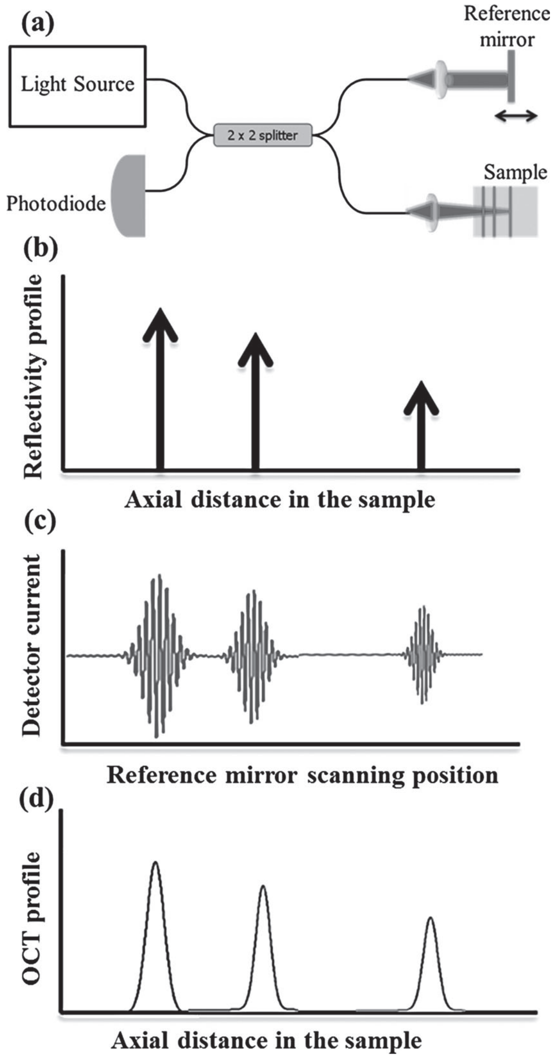


FIGURE 23.2 Illustration of one OCT image profile from a TD-OCT system with different reflectors: (a) a sample with three interfaces, (b) the reflectivity profile of the sample, (c) the output current at the detector, and (d) the OCT image of the reflectivity profile.

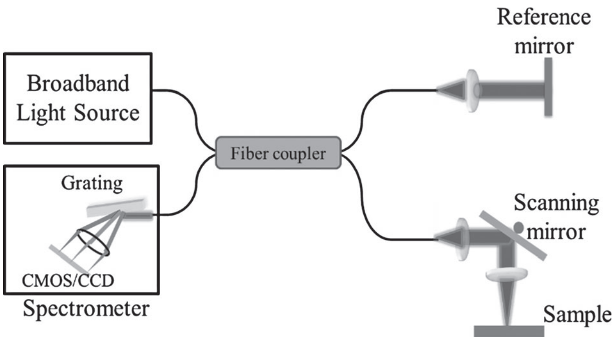


FIGURE 23.3 Schematic layout of a SD-OCT.

the optical wavenumber, by mixing the reference light and the sample light with a fixed group delay. The depth profile of the sample is directly encoded in the interferometric signal in the wavenumber domain, which does not require the mechanical scan of the reference mirror. Studies have shown that FD-OCT has a superior sensitivity advantage over TD-OCT, which is especially beneficial for high-speed imaging (Choma et al. 2003; De Boer et al. 2003; Leitgeb et al. 2003). With the elimination of a moving reference mirror and the advantage of sensitivity, the development of high-speed FD-OCT systems enables comprehensive imaging over larger fields of view and undoubtedly changes the clinical implementation and adoption of OCT (Bouma et al. 2009).

Depending on how the interference signal is recorded, there are two distinct implementations of an FD-OCT system. The first approach, spectral domain OCT (SD-OCT), employs a spectrometer as the detector, which disperses the light spatially and the wavenumber is encoded by the pixel number of the camera in the spectrometer, as schematically shown in Figure 23.3. A spectrometer records the light intensity as a function of wavelength. Typically, an optical grating is used in a spectrometer to disperse the light in space. Different wavelength components of the light are then imaged to a detector array, such as an array of complementary metal oxide semiconductor or charge-coupled devices. The direct output of one A-scan of an SD-OCT system is a number array, which represents the acquired interference signal across different wavelengths and is digitalized by the detector array.

To illustrate the principle of SD-OCT, a perfectly reflecting mirror is assumed to be the sample. If there is a length mismatch of d between the reference arm and the sample arm, it yields an OPL difference of $2d$, as demonstrated in Figure 23.4a. The interference signal is then determined by the following equation:

$$N_g(k) = N_r(k) + N_s(k) + 2\sqrt{N_r(k)N_s(k)}\cos(2kd), \quad (23.1)$$

where k is the wavenumber, $N_g(k)$ is the measured interference spectrum at the spectrometer for wavenumber k, $N_r(k)$ and $N_s(k)$ are the spectra of the reflected light from the reference mirror and sample mirror, respectively. The depth information of the sample is encoded by the spectral oscillations of the interference signal. As shown in Figure 23.4b, the OPL difference of the two arms yields a frequency modulation on the interference spectra, in which the frequency of the modulation increases as the OPL difference increases. For a sample with multiple interfaces, the back-reflected or back-scattered light from each interface interferes with the reflected light from the reference mirror. The output spectrum in this case consists of different modulated frequencies, due to the OPL differences of various interfaces.

Assuming the spectrum is not recorded with evenly spaced wavenumbers, in the post-processing of the SD-OCT signal, interpolations will be needed to resample the signal so the spectrum is sampled linearly in the wavenumber space. Then, the depth profile of the sample can be reconstructed by performing a Fast Fourier Transform (FFT) to the resampled interference signal. The reflectivity profile of the sample along the depth direction, which impacts the visibility of the spectral modulations, will be reconstructed as the amplitude of the FFT signal.

The second implementation of the FD-OCT is swept-source OCT (SS-OCT), which utilizes a wavelength-tunable laser as the light source and a photodiode as the detector; one example of an SS-OCT system is schematically illustrated in Figure 23.5 (Chinn et al. 1997; Lexer et al. 1997). One A-scan is acquired when the tunable light source completes one full sweep over its tuning band. The physics principle of the SS-OCT is similar to the SD-OCT, except that the wavelength of the light is encoded by the time stamps of the sweeping, instead of using the index of pixel number in a detector array.

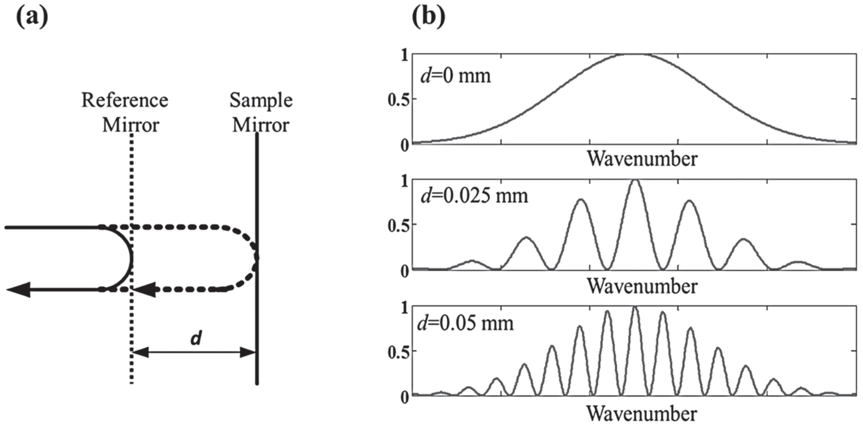


FIGURE 23.4 (a) OPL difference between the reference and sample mirrors. (b) The spectral modulation caused by different OPL differences.

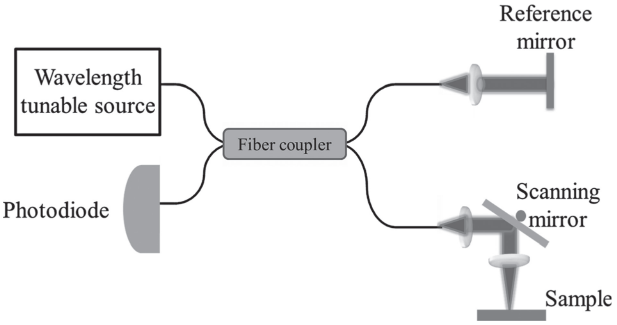


FIGURE 23.5 Schematic layout of an SS-OCT system based on a Michelson interferometer structure.

Compared to SD-OCT, SS-OCT has the advantage of potentially achieving higher imaging speed; an imaging speed of 1.3 million A-scans per second around the 1050 nm wavelength has been reported (Klein et al. 2011). Due to the absence of swept source at shorter wavelengths, most SS-OCT systems operate in the 1050 and 1300 nm wavelength range; with a longer operation wavelength and a limited bandwidth, SS-OCT typically has poor axial resolution compared to that of SD-OCT.

23.2 Broadband Sources for OCT

A broad bandwidth optical source will be the first element that one needs to consider when conceiving an OCT system, since it determines the rest of the optics, the detector, and the system configuration (e.g. spectral domain or swept source-based OCT system). Parameters of the optical sources, such as the spectrum and noise level, impact the OCT system performance significantly. The proper selection of a broadband source is imperative for the OCT instrumentation (Unterhuber et al. 2004). This session provides an overview of the most commonly used broadband sources in OCT systems, including femtosecond lasers, swept sources, supercontinuum sources, and superluminescent diodes (SLDs).

23.2.1 Femtosecond Lasers

A femtosecond laser emits ultrashort light pulses that have a temporal duration in the range of femtoseconds ($1 \text{ fs} = 10^{-15} \text{ s}$). These short laser pulses are attractive for two reasons: a large amount of energy is concentrated within a short time interval and a broadband spectrum can be achieved. For OCT imaging applications, a short pulse with a broadband spectrum is of particularly interesting as it can simultaneously meet the needs for both high brightness (short pulse) and high axial resolution (broadband spectrum) necessary for high-speed acquisition.

The ability of femtosecond lasers to deliver high brightness and broadband spectrum is realized through mode-locking techniques. Mode-locking techniques force different longitudinal modes oscillating in the laser cavity with definite relation between their phases; ultrashort pulses with high intensities can be generated when the phase difference across different longitudinal modes approaches zero. Mode-locking pulses correspond to a frequency comb that represents equally spaced synchronized modes in the frequency domain. Different techniques, including both active- and passive-mode locking, have been developed to generate ultrashort pulses. Among these techniques, Kerr lens mode-locking represents the state-of-the-art method for sub-10-fs pulse generation, in which self-focusing of Gaussian beams occurs due to the strong nonlinear refraction at high peak intensities (Spence et al. 1991). Depending on the choice of the solid-state gain medium, frequently used femtosecond lasers for OCT applications include Ti:Sapphire laser, Cr^{3+}:LiCAF laser, Cr^{4+}:Forsterite laser, and Cr^{4+}:YAG laser.

The Ti:Sapphire laser was first invented in 1982 at the Lincoln Laboratory and uses a crystal of sapphire (Al_2O_3) doped with titanium ions as the lasing medium, which has an extraordinary large optical gain bandwidth (Moulton 1982). A Ti:Sapphire laser is usually pumped with another laser that has a wavelength of 514–532 nm, such as argon-ion lasers (514.5 nm), frequency-doubled Nd:YAG, Nd:YLF, and Nd:YVO lasers (527–532 nm). With the advancement of dispersion control, adapted cavity design, and mirror technology, the Ti:Sapphire laser has been demonstrated to emit pulses with up to 350 nm of spectral bandwidth centred at 800 nm, which yields sub-micron axial resolution of OCT imaging in tissue (Drexler et al. 1999). Ti:Sapphire lasers have been widely used for achieving ultra-high axial resolution in OCT.

With reduced cost and complexity for the pumping laser system compared to that of the Ti:Sapphire laser, the Cr^{3+}:LiCAF laser is pumped directly from a diode and is an alternative broadband laser source for the 800 nm spectral region. An ultrahigh resolution OCT system was demonstrated with a broadband, Kerr-lens mode-locked, diode-pumped Cr^{3+}:LiCAF laser, which has a bandwidth of 89 nm with an output power of 37 mW. The laser enables OCT imaging with 4.5 μm axial resolution in air or 3.4 μm in tissue (Wagenblast et al. 2004).

Since the 1300 nm wavelength enables a larger penetration depth in tissues compared to that of 800 nm, recent efforts also focused on the development of broadband lasers in the 1300 nm spectral region; one example is the development of the Cr^{4+}:Forsterite laser. Bouma et al. demonstrated an OCT imaging system of 6 μm axial resolution with a Cr^{4+}:Forsterite laser that uses a self-phase-modulated Kerr lens mode-locking technique (Bouma et al. 1996). To improve the axial resolution, efforts were made to develop a Cr^{4+}:Forsterite laser with broader bandwidth. A laser spectrum spanning from 1230 to 1580 nm wavelength region has been demonstrated by an all-solid-state Cr:Forsterite laser that has an optical bandwidth of 250 nm (FWHM: Full width at half maximum) (Chudoba et al. 2001; Prasankumar et al. 2002).

The Cr^{4+}:YAG laser is another mode-locked solid-state laser that is capable of emitting ultrashort pulses with a broadband spectrum and high peak power intensities. The Cr^{4+}:YAG laser produces pulses in the spectrum spanning from 1300 to 1600 nm. A bi-directionally pumped Cr^{4+}:YAG laser was developed for OCT with a centre wavelength near 1380 nm and a bandwidth of 222 nm (FWHM) (Hsu et al. 2011).

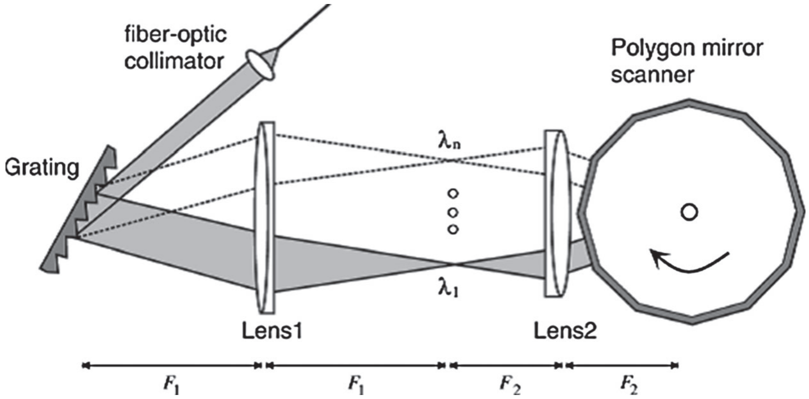


FIGURE 23.6 Schematic layout of a polygon-scanner-based filter (Yun et al. 2003).

23.2.2 Swept Source Lasers

Unlike the broadband femtosecond lasers that emit light with a broad range of spectrum all simultaneously, a swept source emits one wavelength at a time and sweeps across a broad spectral band; it is a common type of light source for OCT. Swept sources are tunable lasers by principle and most swept sources are built from discrete optical components, among which the most important two components are the gain medium and the wavelength-selective filters (Klein and Huber 2017).

The gain mechanisms being employed in swept sources include rare-earth-doped optical amplifiers, stimulated Raman scattering, and semiconductor optical amplifiers (SOAs). Among these gain mechanisms, SOAs are by far the most widely employed gain media for swept sources. Compared to other gain media, SOAs have an ultrashort gain response time (few hundred picoseconds) and also offer broadband optical amplification with gain bandwidths (10 dB) larger than 100 nm, and peak gains of more than 30 dB. This large bandwidth gain of SOAs can be achieved in all popular wavelength windows for OCT, such as centre wavelengths around 1060, 1310, and 1550 nm.

The rapid wavelength-scanning has been achieved by wavelength-selective filters, such as acousto-optic filters, Fabry-Perot filters, and polygon or MEMS mirror-based filters (Yun et al. 2003; Chinn et al. 1997). Figure 23.6 illustrates the schematic layout of a polygon-scanner-based filter, which consists of a diffraction grating, an afocal telescope, and a polygon scanner. As the polygon mirror scans rapidly, only the spectral components that are within a narrow band normal to the front mirror facet of the polygon scanner will be reflected and received by the fibre-optic collimator. The Fabry-Perot filter is another popular implementation of wavelength-selective filters for swept source. Since the Fabry-Perot filter is driven either by electrostatic forces or by a piezo-electric actuator, one of the two end faces of the filter can be moved very rapidly and precisely.

To demonstrate the key parameters of a swept source, the exemplary curves in Figure 23.7a show a linear and unidirectional tuning from short wavelength to long wavelength. The tuning range from short wavelength to long wavelength determines the spectral bandwidth of the swept source, which defines the axial point spread function (PSF) of the OCT system. The sweep period is the duration of the sweep and it determines the A-scan acquisition rate in an SS-OCT. The sweep rate of a swept source for OCT systems usually ranges between 20 and 200 kHz, depending on the applications (Klein and Huber 2017). Figure 23.7b demonstrates the tuning spectrum across one sweep and the instantaneous spectrum (i.e. the emission spectrum at one moment) with a narrow bandwidth. The width of the instantaneous spectrum fundamentally sets the upper limit for the imaging depth of an SS-OCT.

23.2.3 Supercontinuum Sources

Supercontinuum generation is a process in which the laser light is broadened in wavelength (Dudley et al. 2002). The broadening of the spectrum is typically achieved by propagating optical pulses in a highly non-linear medium; for example, one can send optical pulses to propagate through a non-linear optical fibre. As schematically shown in Figure 23.8, supercontinuum can be generated by propagating a femtosecond pulse through a photonic crystal fibre. To generate supercontinuum, a tight confinement of the light energy is important,

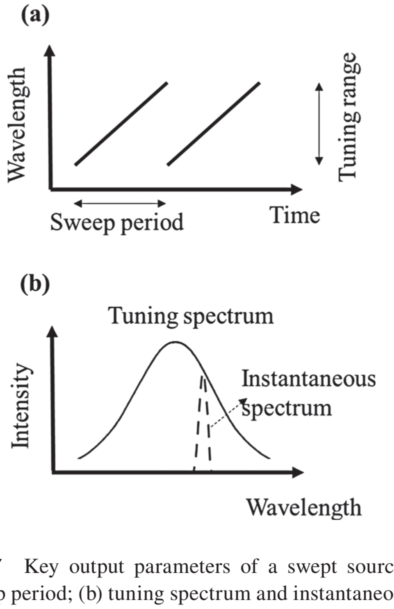


FIGURE 23.7 Key output parameters of a swept source: (a) tuning curve of sweep period; (b) tuning spectrum and instantaneous spectrum.

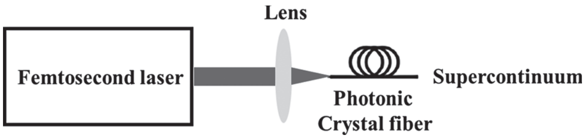


FIGURE 23.8 Schematic layout of supercontinuum generation.

through which the strength of non-linear processes that generates supercontinuum can be enhanced. Both microstructure fibres and tapered fibres with high non-linearity have been explored to generate extremely broad supercontinuum with low-energy femtosecond pulses (Ranka et al. 2000; Birks et al. 2000). The spectral broadening process involves a few non-linear effects, such as self-phase modulation, cross-phase modulation, Raman scattering, and four-wave mixing. These non-linear effects can vary depending on the input pulse and fibre parameters, such as pulse duration, central wavelength of pumping, peak power, dispersion profile, mode area, birefringence property, and fibre length. For example, the spectral broadening will be dominated by self-phase modulation when femtosecond pulses are used as the pump; Raman scattering and four-wave mixing can be important for pumping with picosecond or nanosecond pulses. Supercontinuum generation is also possible with continuous-wave light, when using high-power laser (multi-watt) in long fibres.

A supercontinuum source has a low temporal coherence due to the broad spectrum; the commercially available supercontinuum sources can span an ultra-broad spectral band (e.g. 410 ~ 2400 nm). On the other hand, the spatial coherence of supercontinuum remains high, which does not cause mode dispersion and is important for OCT imaging applications. The low-temporal and high-spatial coherence properties make supercontinuum sources widely adopted in OCT, especially for systems with high axial resolution.

23.2.4 Superluminescent Diodes

The SLD is an edge-emitting semiconductor light source. The principle of SLD light emitting is similar to a laser diode, in which an electronically driven pn-junction becomes optically active when biased in the forward direction, and it generates amplified spontaneous emission over a wide range of wavelengths, which yields low temporal coherence. A schematic layout of a typical SLD is demonstrated in Figure 23.9. The active layer (in which light emits) sandwiched between the p- and n-type cladding layers is formed on an n-type substrate, and voltage is applied across the p-n junction from the electrodes. SLDs are implemented in a waveguide structure with high spatial coherence of the emitted radiation. Due to the simplicity and relatively low cost, SLDs were the major sources in the early development of OCT imaging techniques (Drexler and Fujimoto 2008).

The spectrum of SLDs is determined by the optical gain medium. Most early SLDs in OCT applications had centre wavelengths in the 800 nm regime with a spectral bandwidth of about 20 nm and in the 1300 ~ 1500 nm regimes with spectral bandwidth of about 40 nm. Significant progress in SLD spectral broadening has been achieved after the demonstration of quantum-well SLDs (Chen et al. 1990). Currently, broadband SLDs (up to 100 nm, FWHM) are available at different bands from 680 to 1600 nm, with an output power of a few tens of mW.

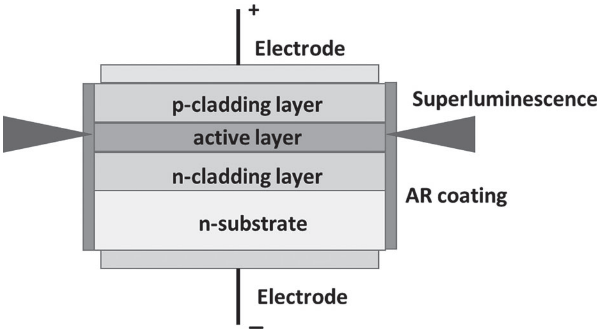


FIGURE 23.9 Schematic layout of a SLD.

23.3 Source Parameters and Their Indications for OCT Performance

The selection of an appropriate optical source is crucial for building an OCT system, since some key OCT performance metrics are determined by the attributes of the optical source, such as resolution and imaging speed. This session will discuss different optical source parameters and their indications for OCT system performance.

23.3.1 Spectrum (Central Wavelength and Bandwidth)

When conceiving an OCT system, the spectrum of a broadband source, including the central wavelength and the bandwidth, is one of the most important parameters to consider. Depending on the applications, the requirement of an optimal central wavelength may be different. For example, most commercially available OCT systems for retinal imaging operate with a central wavelength around 840 nm; however, 840 nm is sub-optimal for imaging the angle structure at the anterior segment of the eye, due to the short penetration depth; a longer wavelength around 1310 nm has been demonstrated to have a better performance for imaging the angle structure with a deeper penetration (Radhakrishnan et al. 2001). Besides the requirement of a specific application, the OCT source spectrum determines the rest of the components of an OCT system, such as the optics and the detector, as well as key performance metrics, such as resolution.

Since OCT is a 3D volumetric imaging modality, the resolution of an OCT system is evaluated in two directions: lateral resolution is perpendicular to the beam propagation direction and axial resolution is along the beam propagation direction. In an OCT system, the lateral and axial resolution are independent due to the different imaging principles; the lateral resolution is determined by the beam spot size at the focus and the axial resolution is determined by the coherence length of the broadband source. Cellular resolution optical coherence microscopy was first reported by Izatt *et al.* and later extended with Gabor domain optical coherence microscopy to a 2×2×2-cubic mm volume in the context of skin and corneal imaging (Izatt 1994; Lee et al. 2011; Tankam et al. 2015).

By using the distance between the central maximum and the first minimum in the Airy disk as a measure, the lateral resolution for circular aperture optical systems is given by

$$\Delta x = 0.61\ \lambda_0 / NA, \tag{23.2}$$

where λ_0 is the central wavelength and NA is the numerical aperture of the objective lens. However, in OCT systems that are implemented with single-mode fibres, the small diameter of the single-mode fibre serves as a pinhole aperture for both illumination and collection of light from the sample; these OCT systems can be treated as a confocal microscope and the lateral resolution can be treated as (Drexler and Fujimoto 2008)

$$\Delta x = 0.37\ \lambda_0 / NA. \tag{23.3}$$

In both Equations (23.2) and (23.3), the lateral resolution scales linearly with the central wavelength; a higher lateral resolution is easier to be achieved with an OCT system that operates at a shorter central wavelength.

The axial "sectioning" ability of an OCT system is based on low-coherence interferometry and the axial resolution (defined as the FWHM of the axial PSF) is determined by the temporal coherence length of the broadband source. For an OCT system with a Gaussian-shaped spectrum, the axial resolution is given by

$$\Delta z = \frac{2\ln 2}{\pi}\frac{\lambda_0^2}{\Delta\lambda}, \tag{23.4}$$

where λ_0 is the centre wavelength and $\Delta\lambda$ is the FWHM bandwidth of the source spectrum. As the spectral bandwidth increases, the axial resolution gets finer; the need for achieving higher axial resolution has been the key driving factor for developing broadband sources. It is also important to note that the axial resolution gets improved quadratically with the decrease in centre wavelength. Figure 23.10 plots the look-up curves of axial resolution vs. spectral bandwidth for three most frequently used central wavelength regimes in OCT (i.e. 800, 1000, and 1300 nm).

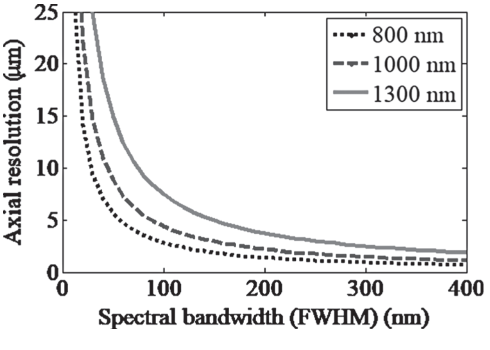


FIGURE 23.10 Axial resolution vs. spectral bandwidth for three different central wavelengths.

For any arbitrary spectral shape, the axial PSF can be evaluated by performing a Fourier transform to the source spectrum (Tripathi et al. 2002). Non-Gaussian spectral shapes will result in sidelobes in the axial PSF, which may generate spurious structures in OCT images and degrade the image quality. For this reason, it is desirable to have a broadband source that has a Gaussian-shaped spectrum for OCT. Different techniques, such as using a programmable spectral processor in hardware or applying a spectral correction curve numerically, have been investigated and demonstrated for shaping the source spectrum to a Gaussian shape (Tripathi et al. 2002; Akcay et al. 2005).

23.3.2 Source Noise

An OCT imaging chain consists of the light emission process of a broadband source, wave propagation through an interferometer, the interaction of light with the sample, and the photon-counting detection process; the acquisition of an OCT signal is a statistical process owing to the presence of noise in different stages of the imaging chain. The noise performance of a broadband source has been an important topic in OCT imaging, since the noise level is closely related to the overall imaging performance through the signal-to-noise ratio (SNR), especially for quantitative imaging tasks. The SNR in OCT is quantified as 20 times the base-10 logarithm of the ratio of the A-scan peak height to the standard deviation of the noise floor. The noise floor standard deviation is quantified at the location of the A-scan peak by blocking the sample arm (Choma et al. 2003). The sensitivity of an OCT system is defined as the SNR of using an ideal reflector as the sample. Different studies have shown that FD-OCT has a significant SNR gain of a few hundredfold compared to that of TD-OCT (De Boer 2003; Leitgeb et al. 2003; Choma et al. 2003). Because of the advantage in SNR, most recent state-of-the-art OCT systems are implemented as FD-OCT. Thus, the source noise will be discussed in the context of FD-OCT.

For both SD-OCT and SS-OCT configurations, the output of one A-scan is a discretized spectrum $\mathbf{N}_g$, which is an array with M elements and M is the number of pixels or time stamps. The digital output at one pixel or time stamp $N_g(\lambda_i, \Delta t)$ is proportional to the number of electrons accumulated in the narrow band centred at λ_i during the detector integration time Δt. Given the source intensity noise as well as the Poisson noise and dark noise of the detector, the variance of the read-out is denoted as $K_{Ng}(\lambda_i, \Delta t)$ and is given as (Huang et al. 2014, 2015)

$$K_{Ng}\left(\lambda_i,\Delta t\right) = C_1\left\langle\left\langle\left\langle N_g\left(\lambda_i,\Delta t\right)\right\rangle\right\rangle\right\rangle^2 + C_2\left\langle\left\langle\left\langle N_g\left(\lambda_i,\Delta t\right)\right\rangle\right\rangle\right\rangle + C_3, \tag{23.5}$$

in which $\left\langle\left\langle\left\langle N_g\left(\lambda_i,\Delta t\right)\right\rangle\right\rangle\right\rangle$ represents the ensembles average of the output over the noise being considered, the coefficients C_1, C_2, and C_3 correspond to the source intensity noise, Poisson noise, and dark noise, respectively. To evaluate the coefficients, a variable neutral density filter can be placed before the input of the detector to adjust the level of light incidence on the detector. The mean and variance of the output at different light incidence levels can be quantified. Measurement examples are

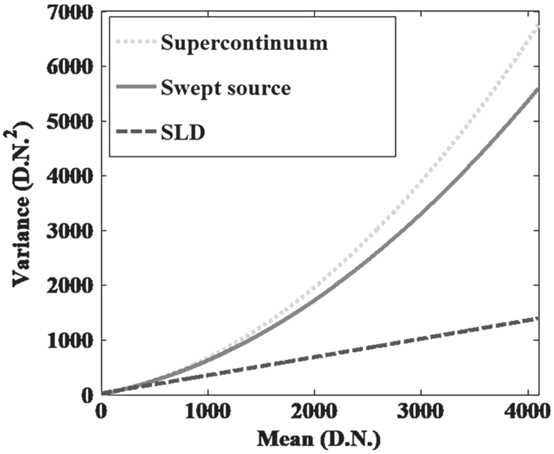


FIGURE 23.11 Noise comparison of different sources (Huang et al. 2015).

shown in Figure 23.11 for a supercontinuum source, an SLD, and a swept source, respectively. Results show that the supercontinuum source is the noisiest, followed by the swept source, and the SLD (Huang et al. 2015).

Another way of quantifying the source noise is to monitor the source power through a photodetector, in which the output of the source is attenuated and incident on a photodetector to avoid saturation or a non-linear response from the detector. The photodetector is then monitored by a RF spectrum analyzer and an oscilloscope. The mean DC of the signal is recorded by the oscilloscope and the corresponding mean square fluctuation is recorded through the RF spectrum analyzer. The intensity noise can be quantified for different detection frequencies (Shin et al. 2010).

Besides the intensity noise, the swept source system suffers from another unique source of noise—the trigger jittering noise. In a SS-OCT system, the swept source sweeps across a broadband spectrum and a start trigger signal is generated to synchronize the detector with the sweeping. The triggering process is caused by the uncertainty of the falling edge of the transistor-transistor logic signal that triggers the detector to start acquisition. The jittering noise results in an uncertainty of the starting point of the spectral interferogram acquisition and randomly changes from cycle to cycle in the wavelength sweeping. The jittering noise can be quantified by using an oscilloscope.

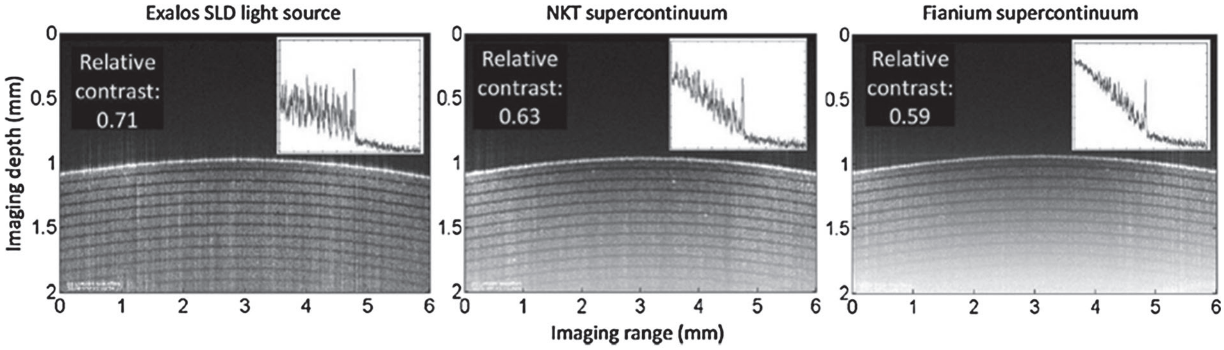
The direct impact of source noise in OCT imaging is the degradation in SNR. An OCT system with a source that has a higher noise level will have a reduced image contrast in one A-scan. Brown *et al.* demonstrated the impact of noise level on image contrast by imaging a roll of tape through OCT systems with different broadband sources. Results show that the SLD has less noise compared to the supercontinuum source and yields images with higher contrast. As shown in Figure 23.12, the relative contrast (measured at the top and bottom of the top layer of tape) is higher for images acquired by a SLD source, compared to that of the image acquired by a supercontinuum source; all sources operate at the same spectral range of 745–925 nm (Brown et al. 2014). When evaluating the impact of source noise, it is also important to note that the evaluation should be considered in the context of the specific imaging task, together with other source parameters such as spectral bandwidth, which may affect the outcome of a task in a significant manner. For example, it was shown that, for thickness estimation, a supercontinuum source, which has the largest spectral bandwidth, yields more precise estimates compared to those of a SLD and a swept source, although the supercontinuum source has the worst intensity noise performance (Huang et al. 2015, 2016).

23.3.3 Instantaneous Linewidth of a Swept Source

In a FD-OCT, different depth information of the sample is encoded by the oscillation frequency of the interference spectrum; a larger depth corresponds to a higher oscillation frequency of the spectrum. According to the Nyquist sampling theory, the maximum imaging depth is given as

$$\pm z_{\max} = \pm \frac{\lambda_0^2}{4 \cdot \delta_s \lambda}, \tag{23.6}$$

where $z_{\max}$ is the maximum one-sided imaging depth, $\delta_s \lambda$ is the sampling interval in wavelength, and λ_0 is the centre wavelength. In an SD-OCT system, the sampling interval in wavelength is determined by the grating and pixel pitch inside



(a) (b) (c)

FIGURE 23.12 OCT images of a tape by using different sources: (a) Exalos 4 SLD, (b) NKT SuperK, and (c) Fianium SC-450-4. Insert graph is an A-scan from the middle of each image going from the bottom of the tape (left of inset) to the top of the tape (right of inset); relative contrast is measured at the top and bottom of the top layer of tape; all sources operate at the spectral range of 745–925 nm. (Adapted from Brown et al. 2014.)

the spectrometer. In an SS-OCT, the sampling interval in wavelength is determined by the sampling rate of the analog-to-digital converter and the sweeping speed of the source. Assuming the source has a linear sweeping in wavelength, Equation (23.6) of the maximum imaging depth can be written as

$$\pm z_{\max} = \pm \frac{\lambda_0^2 f_s}{4\Delta\lambda f_{\text{sweep}}}, \tag{23.7}$$

where f_{sweep} is the sweep rate, f_s is the digitizer sampling rate, and $\Delta\lambda$ is the sweep range of the spectrum (Klein and Huber 2017).

The instantaneous linewidth $\delta_r\lambda$ of the swept source output has a finite spectral bandwidth, which will cause a decrease in interference visibility with an increase in OPL difference. This phenomenon is called the sensitivity roll-off with imaging depth. The roll-off insensitivity may be characterized by the depth at which half of the visibility drops and the depth is given as

$$\pm z_{1/2} = \pm \frac{\ln(2)}{\pi} \frac{\lambda_0^2}{\delta_r\lambda}, \tag{23.8}$$

which is considered as the fundamental limitation of the imaging depth for a given swept source.

23.3.4 Repetition/Sweep Rate of Broadband Sources

The repetition or sweep rate of a broadband source is another parameter one needs to consider when evaluating a source, since they are related with the imaging speed of an OCT system. Currently, femtosecond lasers and supercontinuum sources, which are used for SD-OCT, normally have a repetition rate of multi-MHz; the speed of the SD-OCT is usually determined by the camera speed in the spectrometer, which has a speed up to 200 kHz depending on the pixel number of the line camera. For a swept source-based OCT system, the imaging speed of the OCT is determined by the sweep rate of the swept source, which is on the order of 100 kHz (Klein and Huber 2017). Swept sources with higher sweep rate are available for both research prototypes and commercial products; the commercially available swept sources have achieved a swept rate up to 1.5 MHz.

High-speed imaging is critical for *in vivo* clinical imaging applications and enables the potential for 4D-OCT, i.e. the temporal change of 3D sample structure (Wieser et al. 2014). With applications requiring higher speed in the MHz A-scan regime, the development of higher speed optical sources is one of the future directions for the field of broadband sources in OCT.

23.4 Conclusions

Since the first demonstration of 2D cross-sectional imaging in the early 1990s, OCT has evolved rapidly into a matured volumetric imaging technique with high resolution. One key driving factor to the rapid advancement of the OCT technique is the development of novel broad bandwidth sources, which enable OCT imaging in the visible and near-infrared spectral regime with unprecedented axial resolution. OCT imaging tasks, especially for *in vivo* clinical applications, require both high speed and high sensitivity; the broad bandwidth sources need to have sufficient power output in a single spatial mode in order to meet the demanding requirements of OCT imaging applications. The source candidates that are ideal for OCT instrumentation include femtosecond lasers, swept sources, supercontinuum sources, and SLDs. These sources operate with different physics principles and have very different characteristics.

Femtosecond lasers, which utilize mode-locking techniques to generate ultrashort pulses, provide broadband spectrum that spans almost one octave with relatively low noise. However, these lasers have a complex architecture, which makes them expensive and not feasible for widespread applications. Swept sources, which employ tunable spectral filters to sweep across a broad spectrum, have an advantage for higher speed, and the sensitivity roll-off with depth also performs better due to the narrow instantaneous emission bandwidth (Klein and Huber 2017). One disadvantage of swept sources is the limited tuning spectral range, compared to the spectral bandwidth of femtosecond lasers and supercontinuum sources. SLDs, which generate broadband spectrum by using amplified spontaneous emission, have low noise and are cheap solutions compared to the other broadband sources. The major disadvantages of SLDs are the limited bandwidth and relatively low power. Supercontinuum sources, which achieve spectral broadening by propagating short-pulse lasers in a high non-linear medium, are attractive candidates for OCT systems with an ultra-high axial resolution, due to the spectral bandwidth, available central wavelengths, and output power. However, the relative intensity noise of supercontinuum sources is relatively higher compared to other broadband sources. Different broadband sources have their unique advantages and disadvantages for OCT instrumentation and choosing one appropriate source needs efforts to evaluate the key source parameters, such as central wavelength, spectral bandwidth, power, noise, repetition or sweep rate, and the instantaneous linewidth of a swept source. The evaluation of the parameters should be considered in the context of the specific application and task. To further increase the availability and accessibility of OCT systems, as well as the expansion to more demanding applications, higher speed, lower cost, and more compact broadband sources are highly desirable in the future.

REFERENCES

Akcay, A.C., Clarkson, E. and Rolland, J.P., 2005. Effect of source spectral shape on task-based assessment of detection and resolution in optical coherence tomography. *Applied Optics*, *44*(35), pp. 7573–7580.

Birks, T.A., Wadsworth, W.J. and Russell, P.S.J., 2000. Supercontinuum generation in tapered fibers. *Optics Letters*, *25*(19), pp. 1415–1417.

Bouma, B.E., Tearney, G.J., Bilinsky, I.P., Golubovic, B. and Fujimoto, J.G., 1996. Self-phase-modulated Kerr-lens mode-locked Cr: Forsterite laser source for optical coherence tomography. *Optics Letters*, *21*(22), pp.1839–1841.

Bouma, B.E., Yun, S.H., Vakoc, B.J., Suter, M.J. and Tearney, G.J., 2009. Fourier-domain optical coherence tomography: recent advances toward clinical utility. *Current Opinion in Biotechnology*, *20*(1), pp. 111–118.

Brown, W.J., Kim, S. and Wax, A., 2014. Noise characterization of supercontinuum sources for low-coherence interferometry applications. *JOSA A*, *31*(12), pp. 2703–2710.

Chen, T.R., Eng, L., Zhuang, Y.H., Yariv, A., Kwong, N.S. and Chen, P.C., 1990. Quantum well superluminescent diode with very wide emission spectrum. *Applied Physics Letters*, *56*(14), pp. 1345–1346.

Chinn, S.R., Swanson, E.A. and Fujimoto, J.G., 1997. Optical coherence tomography using a frequency-tunable optical source. *Optics Letters*, *22*(5), pp. 340–342.

Choma, M.A., Sarunic, M.V., Yang, C. and Izatt, J.A., 2003. Sensitivity advantage of swept source and Fourier domain optical coherence tomography. *Optics Express*, *11*(18), pp. 2183–2189.

Chudoba, C., Fujimoto, J.G., Ippen, E.P., Haus, H.A., Morgner, U., Kärtner, F.X., Scheuer, V., Angelow, G. and Tschudi, T., 2001. All-solid-state Cr: Forsterite laser generating 14-fs pulses at 1.3 µm. *Optics Letters*, *26*(5), pp. 292–294.

De Boer, J.F., Cense, B., Park, B.H., Pierce, M.C., Tearney, G.J. and Bouma, B.E., 2003. Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography. *Optics Letters*, *28*(21), pp. 2067–2069.

Drexler, W., Morgner, U., Kärtner, F.X., Pitris, C., Boppart, S.A., Li, X.D., Ippen, E.P. and Fujimoto, J.G., 1999. In vivo ultrahigh-resolution optical coherence tomography. *Optics Letters*, *24*(17), pp. 1221–1223.

Drexler, W. and Fujimoto, J.G. eds., 2008. *Optical Coherence Tomography: Technology and Applications*. Springer Science & Business Media, Berlin.

Dudley, J.M., Provino, L., Grossard, N., Maillotte, H., Windeler, R.S., Eggleton, B.J. and Coen, S., 2002. Supercontinuum generation in air–silica microstructured fibers with nanosecond and femtosecond pulse pumping. *JOSA B*, *19*(4), pp. 765–771.

Fercher, A.F., Hitzenberger, C.K., Kamp, G. and El-Zaiat, S.Y., 1995. Measurement of intraocular distances by backscattering spectral interferometry. *Optics Communications*, *117*(1–2), pp.43–48.

Hsu, K.Y., Jheng, D.Y., Yang, M.H., Lin, Y.S., Huang, K.Y., Liao, Y.H. and Huang, S.L., 2011. Bidirectionally pumped Cr^{4+}: YAG crystal fiber light source for optical coherence tomography. *JOSA B*, *28*(2), pp. 288–292.

Huang, D., Swanson, E.A., Lin, C.P., Schuman, J.S., Stinson, W.G., Chang, W., Hee, M.R., Flotte, T., Gregory, K. and Puliafito, C.A., 1991. Optical coherence tomography. *Science*, 254(5035), pp. 1178–1181.

Huang, J., Yuan, Q., Zhang, B., Xu, K., Tankam, P., Clarkson, E., Kupinski, M.A., Hindman, H.B., Aquavella, J.V., Suleski, T.J. and Rolland, J.P., 2014. Measurement of a multi-layered tear film phantom using optical coherence tomography and statistical decision theory. *Biomedical Optics Express*, *5*(12), pp. 4374–4386.

Huang, J., Yao, J., Cirucci, N., Ivanov, T. and Rolland, J.P., 2015. Performance analysis of optical coherence tomography in the context of a thickness estimation task. *Journal of Biomedical Optics*, *20*(12), p. 121306.

Huang, J., Hindman, H.B. and Rolland, J.P., 2016. In vivo thickness dynamics measurement of tear film lipid and aqueous layers with optical coherence tomography and maximum-likelihood estimation. *Optics Letters*, *41*(9), pp. 1981–1984.

Izatt, J.A., Hee, M.R., Owen, G.M., Swanson, E.A. and Fujimoto, J.G., 1994. Optical coherence microscopy in scattering media. *Optics Letters, 19*(8), pp. 590–592.

Klein, T., Wieser, W., Eigenwillig, C.M., Biedermann, B.R. and Huber, R., 2011. Megahertz OCT for ultrawide-field retinal imaging with a 1050 nm Fourier domain mode-locked laser. *Optics Express*, *19*(4), pp. 3044–3062.

Klein, T. and Huber, R., 2017. High-speed OCT light sources and systems. *Biomedical Optics Express*, *8*(2), pp. 828–859.

Lee, K.S., Thompson, K.P., Meemon, P. and Rolland, J.P., 2011. Cellular resolution optical coherence microscopy with high acquisition speed for in-vivo human skin volumetric imaging. *Optics Letters*, *36*(12), pp. 2221–2223.

Leitgeb, R., Hitzenberger, C.K. and Fercher, A.F., 2003. Performance of fourier domain vs. time domain optical coherence tomography. *Optics Express*, *11*(8), pp. 889–894.

Lexer, F.C.K.H., Hitzenberger, C.K., Fercher, A.F. and Kulhavy, M., 1997. Wavelength-tuning interferometry of intraocular distances. *Applied Optics*, *36*(25), pp. 6548–6553.

Moulton, P., 1982. Ti-doped sapphire: tunable solid-state laser. *Optics News, 8*(6), pp. 9–9.

Prasankumar, R.P., Chudoba, C., Fujimoto, J.G., Mak, P. and Ruane, M.F., 2002. Self-starting mode locking in a Cr: Forsterite laser by use of non-epitaxially-grown semiconductor-doped silica films. *Optics Letters*, *27*(17), pp. 1564–1566.

Radhakrishnan, S., Rollins, A.M., Roth, J.E., Yazdanfar, S., Westphal, V., Bardenstein, D.S. and Izatt, J.A., 2001. Real-time optical coherence tomography of the anterior segment at 1310 nm. *Archives of Ophthalmology*, *119*(8), pp. 1179–1185.

Ranka, J.K., Windeler, R.S. and Stentz, A.J., 2000. Visible continuum generation in air–silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters*, *25*(1), pp. 25–27.

Shin, S., Sharma, U., Tu, H., Jung, W. and Boppart, S.A., 2010. Characterization and analysis of relative intensity noise in broadband optical sources for optical coherence tomography. *IEEE Photonics Technology Letters*, *22*(14), pp. 1057–1059.

Spence, D.E., Kean, P.N. and Sibbett, W., 1991. 60-fsec pulse generation from a self-mode-locked Ti: Sapphire laser. *Optics letters*, *16*(1), pp. 42–44.

Tankam, P., He, Z., Chu, Y.J., Won, J., Canavesi, C., Lepine, T., Hindman, H.B., Topham, D.J., Gain, P., Thuret, G. and Rolland, J.P., 2015. Assessing microstructures of the cornea with Gabor-domain optical coherence microscopy: Pathway for corneal physiology and diseases. *Optics Letters*, *40*(6), pp. 1113–1116.

Tripathi, R., Nassif, N., Nelson, J.S., Park, B.H. and de Boer, J.F., 2002. Spectral shaping for non-Gaussian source spectra in optical coherence tomography. *Optics Letters*, *27*(6), pp. 406–408.

Unterhuber, A., Považay, B., Bizheva, K., Hermann, B., Sattmann, H., Stingl, A., Le, T., Seefeld, M., Menzel, R., Preusser, M. and Budka, H., 2004. Advances in broad bandwidth light sources for ultrahigh resolution optical coherence tomography. *Physics in Medicine & Biology*, *49*(7), p. 1235.

Wagenblast, P.C., Ko, T.H., Fujimoto, J.G., Kaertner, F.X. and Morgner, U., 2004. Ultrahigh-resolution optical coherence tomography with a diode-pumped broadband Cr^{3+}: LiCAF laser. *Optics Express*, *12*(14), pp. 3257–3263.

Wieser, W., Draxinger, W., Klein, T., Karpf, S., Pfeiffer, T. and Huber, R., 2014. High definition live 3D-OCT in vivo: Design and evaluation of a 4D OCT engine with 1 GVoxel/s. *Biomedical Optics Express*, *5*(9), pp. 2963–2977.

Yun, S.H., Boudoux, C., Tearney, G.J. and Bouma, B.E., 2003. High-speed wavelength-swept semiconductor laser with a polygon-scanner-based wavelength filter. *Optics Letters*, *28*(20), pp. 1981–1983.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

24

Laser Applications in Biology and Biotechnology

Sebastian Wachsmann-Hogiu, Alexander J Annala, and Daniel L Farkas

CONTENTS

24.1 Introduction ... 322
24.2 Light Interaction with Matter ... 323
    24.2.1 Absorption of Light ... 323
    24.2.2 Emission of Light ... 323
    24.2.3 Scattering of Light ... 323
    24.2.4 Quantum Confinement—Quantum Dots ... 324
24.3 Laser Probing of Biological Samples ... 324
    24.3.1 Endogenous Molecular Probes ... 324
        24.3.1.1 Endogenous Absorbers ... 324
        24.3.1.2 Endogenous Fluorophores ... 324
    24.3.2 Exogenous Probes ... 325
        24.3.2.1 Small Organic Fluorescent Probes ... 325
        24.3.2.2 Fluorescent Protein-Based Probes ... 325
        24.3.2.3 Quantum Dots ... 326
        24.3.2.4 Raman Probes—Surface-Enhanced Raman Scattering ... 326
    24.3.3 Optical Spectroscopy ... 326
        24.3.3.1 Steady-State Measurements ... 326
        24.3.3.2 Time-Resolved Measurements ... 326
        24.3.3.3 Applications to Tissue Diagnostics ... 327
    24.3.4 Spectral Imaging ... 327
24.4 Laser-Based Imaging: Applications Classified Using the Properties of Light ... 328
    24.4.1 Applications Based on Linear Optical Responses ... 328
        24.4.1.1 Laser Scanning Confocal Fluorescence Microscopy (LSCFM) ... 328
        24.4.1.2 Laser Microscopy Beyond the Diffraction Limit ... 328
    24.4.2 Applications Based on Non-linear Optical Responses ... 330
        24.4.2.1 Multi-photon Laser Scanning Fluorescence Microscopy ... 330
        24.4.2.2 Imaging Techniques Exploiting Second Harmonic Generation (SHG) ... 330
        24.4.2.3 Imaging Techniques Employing Coherent Anti-stokes Raman Scattering (CARS) ... 331
        24.4.2.4 Surface-Enhanced Raman Scattering ... 331
    24.4.3 Spectral-Based Applications—Imaging Beyond the Visible Region ... 331
        24.4.3.1 Imaging with Soft X-rays in the ‘Water-Window’ Region ... 331
        24.4.3.2 Imaging with Lasers Operating in the ‘Water-Window’ Region ... 332
24.5 Laser-Based Imaging: Applications to Different Levels of Biological Organization ... 333
    24.5.1 Single-Molecule Detection ... 333
    24.5.2 Imaging at Sub-cellular Level ... 333
    24.5.3 Imaging of Organelles and Bacteria ... 333
    24.5.4 Imaging at Cellular Level ... 333
    24.5.5 Imaging in Tissue ... 334
24.6 Lasers for Micromanipulation of Biological Samples ... 334
    24.6.1 Opto-Mechanical Manipulation of Biological Samples ... 334
        24.6.1.1 Optical Tweezers ... 334
        24.6.1.2 Perforation (Optoporation) ... 335
        24.6.1.3 Nanosurgery (Laser Scalpels) ... 335
        24.6.1.4 Tissue Welding and Soldering ... 335
    24.6.2 Physico-Chemical Manipulation of Biological Samples ... 335
        24.6.2.1 Pump-Probe Technique ... 335

24.6.2.2 Laser-Activated Processes 336
24.6.3 Fluorescence Correlation Spectroscopy 336
24.6.4 Flow Cytometry 336
24.7 Emerging Applications of Lasers in Biotechnology 337
24.7.1 DNA Micro-arrays and Genomics 337
24.7.2 Protein Micro-arrays and Proteomics 337
24.7.3 Bio-Cavity Lasers 337
References 338

24.1 Introduction

The laser is, due to the special nature and quality of laser light, a vital tool in nearly every aspect of everyday life including communications, entertainment, manufacturing and medicine. An identifiable icon of technology, the laser has become indispensable for science. Many important discoveries of the past half-century are related to the use of lasers. The field of applications is growing rapidly, as performance parameters improve and laser technology evolves towards smaller lasers, higher power and shorter pulses. We will survey here the very important applications of lasers in biology and biotechnology.

The fundamental nature of light was a topic of hot debate among scientists for a very long time but its usefulness for observation—including scientific research—was never questioned. In fact, every one of the rich properties of light has been used for experimental investigation, spanning more in temporal and spatial domains than any other tool. Figure 24.1 shows the wavelengths of some of the most commonly used lasers in biological applications.

The first laser, built in 1960 by Theodore Maiman, used a synthetic ruby rod that emitted intense pulses of coherent red light at 694 nm in a tightly concentrated beam, quite typical of the characteristics displayed by many current lasers. Technology, performance and applications have greatly evolved since, with the discovery of different active media making laser light available from the UV to the infrared region of the electromagnetic spectrum.

Ultrashort pulsed lasers, from free-running lasers to Q-switched lasers (nanosecond pulse duration) and mode-locked lasers (picosecond and femtosecond pulses), have revolutionized entire fields of physics and chemistry and extended the horizons of what is accessible to study in biology. With femtosecond pulses it has been possible to see how atoms in a molecule move during a chemical reaction—a remarkable achievement recognized by a Nobel prize for A Zewail in 1999. Photochemical reactions like those in photosynthesis, vision or photodynamic therapy can be studied to gain information about how light energy is converted into chemical energy or to learn how to use light in cancer detection and treatment.

For many purposes in photobiology, it is necessary to mark or probe selected regions of cells or tissue by introducing compounds that will interact with light they absorb in a predetermined way. Such compounds are known as probes and may be either added artificially to the sample (*exogenous probes*) or be naturally present in the cell or tissue sample (*endogenous probes*). To understand how such markers work, we need to consider the ways in which light interacts with molecules.

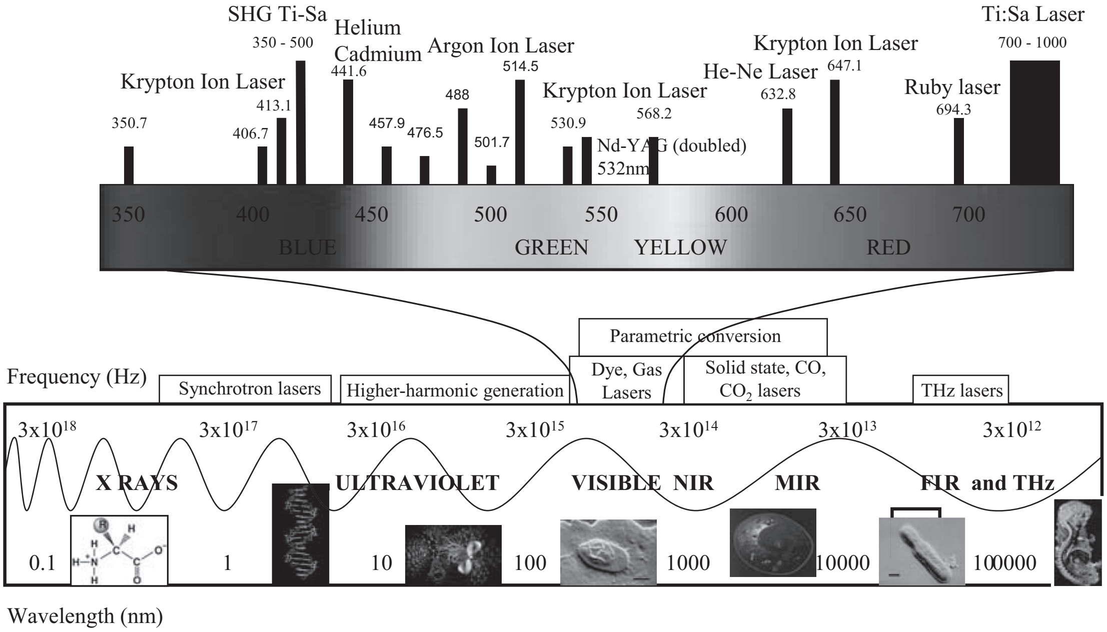

SHG Ti-Sa
350 - 500
Helium Cadmium
441.6
Argon Ion Laser
514.5
488
Krypton Ion Laser
413.1
350.7
406.7
457.9
476.5
501.7
Krypton Ion Laser
530.9
568.2
Nd-YAG (doubled)
532nm
He-Ne Laser
632.8
Krypton Ion Laser
647.1
Ruby laser
694.3
Ti:Sa Laser
700 - 1000
350
400
450
500
550
600
650
700
BLUE
GREEN
YELLOW
RED
Parametric conversion
Frequency (Hz)
Synchrotron lasers
Higher-harmonic generation
Dye, Gas Lasers
Solid state, CO, CO2 lasers
THz lasers
3x10^18
3x10^17
3x10^16
3x10^15
3x10^14
3x10^13
3x10^12
X RAYS
ULTRAVIOLET
VISIBLE
NIR
MIR
FIR and THz
0.1
1
10
100
1000
10000
100000
Wavelength (nm)

FIGURE 24.1 The electromagnetic spectrum showing commonly used laser lines.

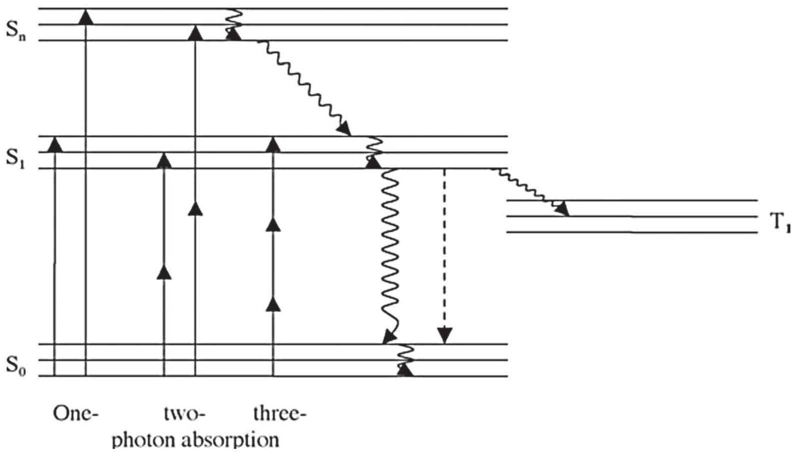


FIGURE 24.2 Absorption and emission processes in a dye molecule. S_0 is the electronic ground state, S_1 is the first electronic excited state, S_n is the *n*th excited electronic state, and T_1 is a triplet state. One-, two- and three-photon absorption is indicated by straight, full lines, emission by dashed lines. The non-radiative processes and intersystem crossing are indicated by curvy lines.

24.2 Light Interaction with Matter

24.2.1 Absorption of Light

Absorption is a transfer of energy from the electromagnetic wave to the atoms or molecules of the medium. The energy transferred can either excite electrons to higher energy states, if this energy matches the difference between the electronic energy levels of the atoms or molecules, or excite vibrations or rotations in the molecules. The wavelengths of light that can excite these energy states depend on the energy-level structures and, therefore, on the types of atoms and molecules contained in the medium. The spectrum of the light after passing through a medium appears to have certain wavelengths removed because they have been absorbed. This is called an absorption spectrum.

Molecules can be excited by the absorption of a single photon of wavelength λ which promotes the molecule from an unexcited energy level of energy E_0 belonging to the electronic ground state S_0 thereby promoting the molecule to an excited level of energy E_1 belonging to the first electronic state S_1, provided the frequency v of the photon is resonant with the energy gap $E_1 - E_0$:

$$E_1 - E_0 = hv = hc/\lambda \qquad (24.1)$$

where h is Planck's constant and c is the velocity of light. However, much weaker absorption on the same transition can also occur by the simultaneous absorption of two photons at exactly half the frequency, i.e. at the frequency $v/2$ corresponding to the wavelength 2λ. Because this is a non-linear effect, it requires very high intensities in order to proceed at a reasonable rate. This is why this type of effect only became of practical importance with the introduction of laser technology. An even weaker effect but one which is readily observable with even higher laser intensities is the simultaneous absorption of three or more photons of appropriate energy. For an absorption transition that proceeds as a single photon absorption process at a wavelength λ lying in the ultraviolet region of the spectrum at, for example, 340 nm, the two-photon absorption will occur by excitation in the red region at 680 nm and the three-photon absorption in the NIR region at 1020 nm (Figure 24.2).

24.2.2 Emission of Light

Molecules can lose their excitation energy in a number of ways. For example, collisions return them to lower energies without the emission of radiation. However, for certain types of molecules in the liquid phase, they may emit radiation by spontaneous emission on the same transition at the same wavelength as that of single-photon absorption from S_0 to S_1. This process is called fluorescence and occurs with typical spontaneous emission lifetimes of a few nanoseconds.[1]

Those molecules (fluorescent dyes, generally polyaromatic hydrocarbons or heterocycles) which show efficient emission of fluorescence are called fluorophores, and are still the most commonly used fluorescence markers (although the advent of quantum dots may change this in the future—see Section 24.3.2.3). To fluoresce, the dye needs to be excited first either by absorption of a photon or by chemical means.

24.2.3 Scattering of Light

Scattering is the redirection of light caused by the interaction of light with matter. The scattered electromagnetic radiation may have the same wavelength (energy) as the incident radiation (Rayleigh scattering) or a different wavelength (as in Raman scattering) and it may have a different polarization.

The Raman effect (a much weaker effect than Rayleigh scattering) describes inelastic scattering of the incident light from the sample: the scattered light is shifted to higher and lower frequencies compared with the incident light by an amount of energy characteristic for the molecular vibrations. Given its specificity to molecular vibrations, this is a powerful technique for molecular biology. Laser light sources have increased the sensitivity and the usefulness of the method.

---

[1] There is also another process known as phosphorescence which involves excitation to the triplet state T_1 from which the excitation decays by emission at a longer wavelength and much slower lifetimes, in the order of a few µs. This process is of importance in understanding the dynamics of dye lasers but will not concern us in the present discussion.

The Raman spectrum can be measured non-invasively under ambient conditions and without special sample preparation. It can provide information about the molecular make-up of the sample and, thus, help elucidate the chemical structure and composition of very complex systems. Raman spectroscopy can also be used for ultra-sensitive trace detection.

24.2.4 Quantum Confinement—Quantum Dots

Light can also be emitted from semiconductors as a result of recombination of electron–hole pairs that, in turn, are created by exciting the semiconductor either electrically or by shining light on it. The energy and, therefore, the wavelength of the emitted light are governed by the composition of the semiconductor material and for bulk samples of the most common semiconductors lies in the infrared region of the electromagnetic spectrum. If, however, the physical size of the semiconductor sample is made very small (to dimensions much smaller than the natural radius of an electron–hole pair), the energy levels of the electron in its potential well shift upwards and spread apart. In the case of particles with dimensions of the order of a few nanometres, this leads to a shift in emission towards frequencies which lie in the visible region of the spectrum. The spectral region in which light emission occurs can, therefore, be tailored by creating particles of the appropriate size and composition. Such particles are called quantum dots. They exhibit a high-fluorescence quantum yield, are non-photobleaching and have narrow, symmetric emission spectra and broad absorption spectra, making them a close approximation to the ideal fluorophore (see also Section 24.3.2.3).

24.3 Laser Probing of Biological Samples

The lasers commonly used to probe biological samples are medium-intensity monochromatic light sources such as argon-ion lasers, dye lasers and diode-pumped solid-state lasers. They are applied in a variety of techniques including scanning confocal microscopy, spectral microscopy, lifetime imaging studies, photobleaching recovery and total internal reflection fluorescence. They have also been utilized, although less frequently, in conventional wide-field fluorescence investigations.

The optical signal detected upon the interaction of laser light with biological samples contains information not only about the biochemical content but also about the structure of the investigated sample and physico-chemical processes in which they are involved. Measurements of absorption, fluorescence, phosphorescence, Raman and Rayleigh scattering require the presence of endogenous or exogenous probes in biological samples. The molecular probes should be intrinsically non-toxic and not generate phototoxic products, like triplet oxygen, under photoexcitation. In addition, they should not hinder/perturb the biology and chemistry to be investigated.

24.3.1 Endogenous Molecular Probes

24.3.1.1 Endogenous Absorbers

Two main classes of endogenous chromophores, based on the property that allows them to be used as reporters for structures and/or events of biological significance, can be distinguished: absorbers (chromophores that only absorb light) and fluorophores (chromophores that both absorb and re-emit light).

The compounds primarily responsible for absorption of light in tissue are oxy- and deoxyhaemoglobin, melanin, myoglobin, rhodopsin and water. Functional events can be studied by monitoring optical changes in endogenous pigments (e.g. rhodopsin, carotenoids in the retina). Since tissue scattering predominates over absorption in the NIR range (650–1000 nm), backscattering measurements at these wavelengths are used to characterize the tissue structure (surface, layers/interface, thickness, inhomogeneities). In addition, since the changes in the haemoglobin absorption spectra are indicative of the physiological status of tissue, oxygenation measurements can provide diagnostic information on tissue oxygenation and oxygen consumption [1,2].

24.3.1.2 Endogenous Fluorophores

Photosynthetic systems, having evolved to harvest and convert (sun)light energy effectively, contain a large array of pigments and pigment–protein complexes, with strong absorption-emission characteristics. Lasers have proven extremely useful in the study of photosynthesis. In particular, primary photophysical events of importance cannot really be investigated without fast lasers and this field of research has flourished in parallel with the developments in laser technology [3–5].

In certain organisms, changes in dominant constitutive pigments (such as bacteriorhodopsin in *Halobacterium halobium*) constitute the best avenue for study, especially for fast events [6].

In largely non-pigmented tissue, fluorescence commonly originates from several endogenous fluorophores: aromatic amino acids, structural proteins, enzyme co-factors involved in cellular metabolism, lipopigments and porphyrins.

The main contributors to protein fluorescence are the aromatic amino acids tryptophan, tyrosine, and phenylalanine [7]. They exhibit high absorbance at wavelengths below 300 nm and emission peaks in the 260–400 nm range. Quantum yields vary between 0.02 (phenylalanine) and 0.14 (tryptophan), with lifetimes ranging from 1 to 6 ns, depending upon local environment.

Tissue fluorescence is associated with the structural or extracellular matrix proteins elastin and collagen. Based upon chromatographic techniques, the fluorescence of the collagen family of molecules has been attributed to various amino-acids and amino-acid cross links. Several fluorescent cross links have been identified, including desmosine, isodesmosine (and the tri-amino pyridinium derivative for elastin), and pyridinoline and pentosidine for collagen. Both elastin and collagen have a broad range of absorption spectra from below 300 nm to above 400 nm.

The emission spectra vary with the wavelengths of excitation. For instance, collagen peak emission ranges from 310 nm (with excitation at 280 nm) to 520 nm (with excitation at

450 nm). Several components of collagen were found to absorb in the far-UV (aromatic amino acids: phenylalanine, tyrosine), the mid-UV (carbonyl components, aldol condensation cross links, keto-amine cross links) and the near-UV (pyridinoline). Consequently, the fluorescence spectra reflect the emission of particular cross links that are most efficiently excited at a certain UV range. Moreover, the cross links were found to be specie-dependent. Differences in fluorescence emission were identified between collagen in bovine tendons, cartilage, bone, dentine, vitreous humor and collagen in rat tail tendon, skin and cornea. Fluorescence lifetimes also vary with the type of collagen type and/or source [8].

Fluorophores involved in cellular metabolism include two enzymes co-factors, nicotinamide adenine dinucleotide (NADH) and flavin adenine dinucleotide (FAD). Only the reduced form of NADH and the oxidized form of FAD are fluorescent, as reviewed by Lakowicz [7]. NADH absorbs in the 300–390 nm wavelength range and emits in the 400–600 nm range (lifetime ~0.4 ns in an aqueous buffer). Upon binding to protein, the quantum yield of NADH is reported to increase about four-fold and the lifetime increases to 1.2 ns. FAD has absorption in the visible range (400–500 nm) and emits around 525 nm.

Lipopigments in tissue that are the end products of lipid metabolism have been associated with pathological processes (atherosclerosis, retinal degeneration) and aging [9]. Fluorescent lipopigments include ceroid and lipofuscin. Both absorb light in the UV range and have an intense fluorescence emission in the 450–650 nm range.

Pathological or physiological changes in tissues result in associated molecular transformations. Derangements can be detected by changes in the relative contribution of each fluorescent constituent to the overall fluorescence emission. Owing to its correlation with physiological and biochemical characteristics, autofluorescence represents an important measure of cell and tissue characteristics and changes, thus enabling their investigation in a native state (see, e.g. Mayinger [10], Millot [11] and Shibuki [12]).

24.3.2 Exogenous Probes

Many of the laser techniques used in biology and biotechnology need markers or exogenous probes (enhancers) that specifically bind to the bioparticle of interest in order to reveal information about the sample. An exogenous probe is an entity (e.g. a dye molecule, a semiconductor or metallic nanoparticle) designed to localize within a specific region of a biological specimen and respond to a specific stimulus. A fluorescent probe should have a high fluorescence yield, photostability and yet retain the critical parameters of the unlabelled bioparticle, including selective binding to a receptor, activation or inhibition of a particular enzyme or the ability to incorporate into a biological membrane.

24.3.2.1 Small Organic Fluorescent Probes

Organic dyes are small molecules that can absorb and subsequently emit light with a high quantum yield. These probes can bind either (i) directly to chemical reactive groups in the bioparticles for which they have a chemical affinity or (ii) indirectly using the high affinity and specificity of avidin–biotin and anti-body–hapten interactions.

The dyes of the first class have reactive groups like amines, thiols, alcohols, etc. capable of creating strong chemical bonds between the dye and the biomolecule. They are generally used to prepare fluorescent proteins, peptides, ligands, synthetic oligonucleotides and other biomolecules for probing biological structure, function and interactions. Some of the areas of interest are immunochemistry, fluorescence *in situ* hybridization (FISH), cell tracing, etc. They are helpful in analyzing the topography of proteins in biological membranes, determining distances within a protein or between a protein and a ligand using excited-state energy transfer (fluorescence resonance energy transfer) and subsequent changes in protein conformation using environment-sensitive probes.

The second class of dyes is based on avidin–biotin and anti-body–hapten specific interactions. The avidin or anti-hapten anti-body conjugate needs first to be labelled with a fluorophore. The molecule of interest (such as an anti-body, lectin, drug, polynucleotide, polysaccharide or receptor ligand) is then tagged with biotin or hapten. The specific interaction of avidin–biotin and anti-body–hapten enables the biotinylized or haptenylized molecule to be monitored—this would otherwise be very difficult to detect. This method is used to probe complex solutions, cells and tissues, as well as protein and nucleic acid blots and arrays. It can also be extended to simultaneous multi-colour detection of multiple targets in systems of interest.

Other small fluorescent probes are calcium indicators and caged probes. Fluorescent calcium indicators include Fura-2, Indo-1 and Fluo-3. They continue to be the most sensitive and popular tools for monitoring the concentration of cytosolic free calcium in living cells. One of the problems associated with these indicators is their tendency to accumulate in organelles and leak out of cells.

Caged dyes are molecules that can be turned on by brief exposure to UV (see also 24.6.2.2(c)). They are important tools for delivering physiological stimuli by naturally active biomolecules with spatial and temporal precision that far exceeds that of micro-injection or perfusion.

One of the best references for fluorescent probes can be found on the website of Molecular Probes, Inc.: http://www.probes.com.

24.3.2.2 Fluorescent Protein-Based Probes

The primary advantages of fluorescent protein-based probes are that they can be designed to respond to a much greater variety of biological events and signals, may be targeted to sub-cellular compartments, introduced into a wider variety of tissues and intact organisms and they very rarely cause photodynamic toxicity. The use of fluorescent proteins as exogenous probes began with the discovery that the jellyfish *Aequorea victoria* produces a naturally fluorescent protein known as green fluorescent protein (GFP). In jellyfish, the protein serves probably as a Foerster-transfer-based wavelength conversion mechanism for chemi-luminescence. The gene for this protein has been cloned and transfected into other organisms and revolutionized many areas of biological fluorescence microscopy. By combining the gene sequence for a certain endogenous protein with the sequence for one of the GFPs, variants of a chimeric fusion-protein can be produced. Such a construct often still performs the biological role of the original protein while permitting the observation of the spatio-temporal concentration dynamics of that protein. Other possibilities include the specific labelling of certain cell types by putting the GFP under the control of gene control sequences (promoters) that are selectively active only in these cells or only in certain developmental or physiological states.

There are presently several variants of GFP isolated from other organisms or created using a combination of mutagenesis and structure-based protein engineering. They now provide spectrally separable emission colours, with some of these variants having emission spectra that are no longer ‘green’ [13]. Yellow, blue and red versions of GFP are thus called YFP, BFP and RFP, respectively. Furthermore, in analogy to synthetic fluorescent indicators, variants of GFPs with strong sensitivity to their chemical environment have been protein-engineered [14].

Interesting applications come with the development of a free-calcium probe based on GFP. Two colour variants of GFP have been engineered into a single chimeric molecule that contains calmodulin, a calcium-binding protein [14]. The chimeric molecule has been so engineered that, upon binding calcium, the fluorescent peptides undergo fluorescence resonance energy transfer. The emission from the shorter wavelength fluorescent peptide is quenched and energy is transferred to excite the longer wavelength peptide, thus changing the ratio of the two emitted wavelengths. The result is a calcium indicator for all cells within the organism for which the gene promoter is active. An excellent review on protein probes is by Zhang et al. [15].

24.3.2.3 Quantum Dots

Quantum dots (also called nanocrystals) absorb light and then quickly re-emit light with a different colour (see Section 24.2.4). They are nanometre-sized semiconductor particles that exhibit a high-fluorescence quantum yield, do not photobleach and have broad absorption spectra with narrow, symmetric emission spectra. The emission spectra can be ‘tuned’ to any chosen wavelength by simply changing the size of the quantum dot (smaller dots emit in blue, larger dots in red). The broad absorption enables simultaneously excitation of multiple different-size quantum dots with the same laser wavelength. These properties make them a close approximation to the ideal fluorophore.

Quantum dots need to undergo special treatment to be applicable in biological systems. They should be made soluble in water and non-toxic for biologic material [16]. Two-photon laser scanning microscopy images of nanocrystals have been obtained, and because two-photon cross sections are expected to be huge compared to molecular fluorophores, nanocrystals may become the method of choice for the tracking of single fluorophores in biological tissue.

24.3.2.4 Raman Probes—Surface-Enhanced Raman Scattering

The Raman signal observed in biological samples is extremely weak and often difficult to measure. However, it is possible to enhance the signal by attaching the molecule of interest to metallic nanostructures. This phenomenon, called surface-enhanced Raman scattering (SERS), can produce enhancement of Raman scattering of up to 14 orders of magnitude (see Section 24.4.2.4).

SERS is observed primarily for analytes adsorbed onto transition metals of group 11 (Au, Ag, Cu) or alkali (Li, Na, K) metal surfaces. Theoretically, any metal would be capable of exhibiting surface enhancement but the noble and alkali metals provide the strongest enhancement [17]. Metals such as Pd or Pt exhibit enhancements of about two- to three-fold for excitation in the near ultraviolet. An important factor in enhancement is the nanoscopic roughness of the metallic surface. A wide variety of substrates has been found to exhibit SERS: electrochemically modified electrodes [18], colloids [19], island films [20] and, later, regular particle arrays [21–23]. In general, due to its chemical inactivity, gold is the most suitable metal for incorporation inside biological entities like living cells. Gold colloidal clusters have very good SERS enhancement factors when NIR excitation is applied.

24.3.3 Optical Spectroscopy

Most generally, spectroscopy studies the way in which a particular property of a system is distributed over its components. Scientific/optical spectroscopy is concerned with analyzing the emission and absorption of electromagnetic energy by atoms and molecules. This can be done based on parameters such as wavelength and time. The probing of material with (electromagnetic) waves of specific frequency and wavelength and the study of the response of the material to such probes are of special relevance to biological applications. Spectroscopy is employed in a wide range of biological research areas, such as biochemistry and toxicology, for the identification of metabolites and other compounds of biological significance. Of these, fluorescence is the most widely utilized with measurements that can be categorized as either static (steady-state or time-integrated) or dynamic (time-resolved).

24.3.3.1 Steady-State Measurements

Steady-state techniques provide an ‘integrated’ spectrum over time that gives information about fluorescence emission intensity and spectral distribution. In general, the fluorescence of biological samples is induced by a continuous light source and then recorded with the help of a detector array such as an optical multi-channel analyzer (OMA) or by a scanning monochromator connected to a photomultiplier. Although convenient, scanning the spectrum by means of a monochromator allows only the light of a particular wavelength to be recorded at a specific time. This drawback is compensated by the superior sensitivity of photodetectors and by their linear response, in contrast to the diode array used by an OMA system. Due to the relative simplicity of the instrumentation, measurements based upon steady-state methods are the most common for biological tissue research.

24.3.3.2 Time-Resolved Measurements

Time-resolved techniques record the changes (decay) of fluorescence emission intensity as a function of time and, thus, provide deeper insight into the molecular species of the sample (e.g. the number of fluorescent species and their contribution to the overall emission), quenching processes and/or changes in the local environment. Two methods of time-resolved measurements are typically used: time domain and frequency domain. For time-domain measurements, the sample fluorescence is induced by a short pulse of light, of the order of nanoseconds or shorter, and the emission is recorded with a fast and sensitive photodetector. The most common technique for time-domain measurements is time-correlated single-photon counting. For the frequency domain, the sample is excited with intensity-modulated light. Typically, the intensity of the excitation light is varied at a high frequency. These methods are described in detail by Lakowicz [7,24,25].

24.3.3.3 Applications to Tissue Diagnostics

Pertinent reviews in the field of laser-induced fluorescence spectroscopy (LIFS) for tissue diagnostic applications are presented by Wagnieres et al. [26], Bigio and Mourant [27], Andersson-Engels et al. [28] and Richards-Kortum [29]. While steady-state LIFS techniques have been extensively investigated and are currently routinely used in research clinics, the time resolved-LIFS techniques for tumour diagnostics have been explored only by a few research groups. Both time- and frequency-domain time-resolved instrumentation have been employed.

Non-linear (multi-photon) spectroscopy of tissues has been reviewed by Birch [30].

24.3.4 Spectral Imaging

Light, upon coming in contact with an object, can be transmitted, reflected/scattered, absorbed and re-emitted. Spectroscopy measures these phenomena in the temporal or frequency domains in order to determine important properties of the object being probed. Depending on the energy of the light and the nature of the object, some of these interactions, such as absorption by particular chemical bonds, can reflect the presence and quantity of certain well-understood constituents. The addition of *imaging* to spectroscopy can trace its origins to pioneering satellite-based remote-sensing, developed on the premise that information present in the optical properties of natural and man-made objects could be used to detect and monitor these entities from a distance. Similar premises are at the basis of non-contact studies by biological spectral imaging, particularly with 'optical biopsy' [31,32].

Human colour vision is a form of imaging spectroscopy, by which we determine the intensity and proportion of wavelengths present in our environment. Spectral imaging improves on the eye in that it can break up the light content of an image not just into red, green and blue but into an arbitrarily large number of wavelength classes. Furthermore, it can extend the range to include the invisible ultraviolet and infrared regions of the spectrum denied to the unaided eye. This type of imaging is usually known as hyperspectral. The result of (hyper) spectral imaging is a data set, known as a data cube, in which spectral information is present at every picture element (pixel) of a digitally acquired image [33].

The information conveyed by colour (or spectral features in general) varies. In some instances, colour reflects the presence of a single, important chromophore which can be detected and quantified by virtue of its absorbing certain wavelengths. Haemoglobin, chlorophyll, bilirubin, etc., are such molecules. Furthermore, some substances will reflect changes in their functional status with absorbance shifts, the obvious example being the colour shift involved in going from oxy- to deoxyhaemoglobin (the basis for pulse-oximetry used clinically to monitor oxygenation). In other, more complex cases, colour can function as a signature, signifying an identity, even when we do not understand all the physical bases for the perceived colour composition. In contrast to intrinsically coloured objects, when histology or pathology samples are examined in transmission under a microscope, they generally appear almost clear and, especially when single cells are examined, will be virtually invisible unless optical techniques such as phase or differential interference contrast microscopy are used. In other words, in these cases, the specimens are optically weak and there is almost no usable spectral information in the visible range. Therefore, staining is required before viewing. The most commonly used stains provide intense colour signals but the relationship between the observed colours and the cellular constituents is obscure at best and variable at worst.

Finally, and most importantly, exogenous, spectrally defined chromophores or fluorophores (see Section 24.3.2) can be coupled to agents that bind to cellular components with great specificity, in such techniques as ion-sensitive signalling [34], immunohistochemistry [35,36], immunofluorescence [37], and FISH [38,39]. In these cases, the colours usually bear only an arbitrary connection to the molecular structures they are highlighting by being coupled to them, although in some cases spectral shifts and/or intensity changes can be used to detect intra- and inter-molecular interactions, and this includes live cell and tissue imaging [36,40,41]. In such imaging experiments, lasers have been used very successfully as light sources, particularly since their narrow, well-defined wavelengths facilitate emission-based spectral separation.

A large and growing number of spectral imaging techniques exist, including Fourier—scanning and non-scanning—methods [42,43], band-sequential methods such as liquid crystal and acousto-optic tunable filters [44–47] and more traditional ones, e.g. those which are monochromator or grating-based. This means spectral discrimination is no longer the main challenge. A number of new applications of spectral imaging in biology and medicine have been described [33,48–51], with spectral karyotyping [52], spectral pathology [31,32], gene expression [53] and physiological measurements such as blood oxygenation [1,2] appearing particularly exciting.

Analysis of spectral imaging data has engendered a large body of theoretical and practical work but much of it is still mainly directed towards remote sensing applications [42], and is just beginning to find its way into biological research applications [44,54]. The next step, then, for biological and medical spectral imaging will be to develop robust software that will permit quantitative, intuitive analysis. Any sophisticated analysis should also account for the fact that, unlike satellite reconnaissance, which relies on natural (mostly reflected) light, biological spectral imaging can be focusing on specific signals, such as fluorescence, elicited by laser illumination sources with the usable wavelengths shown in Figure 24.1.

24.4 Laser-Based Imaging: Applications Classified Using the Properties of Light

24.4.1 Applications Based on Linear Optical Responses

24.4.1.1 Laser Scanning Confocal Fluorescence Microscopy (LSCFM)

In a conventional microscopy, not only is the plane of focus illuminated but much of the specimen above and below it is also illuminated at the same time. This results in an out-of-focus blur from the areas above and below the plane of interest, leading to a reduction in image contrast and a decrease in resolution.

In LSCFM, the laser light is expanded to make optimal use of the optics in the objective. This beam is focused to a spot down to 0.25 µm in diameter and 0.5 µm deep (depending mostly on the excitation wavelength) and is scanned across the fluorescent sample by an *x–y* deflection mechanism. The excitation and collection is, thus, sequential in nature. The mixture of emitted fluorescence, scattered and reflected light is captured by the same objective and is focused onto a photodetector via a dichroic mirror (beamsplitter). A confocal aperture (pinhole) is placed in front of the photodetector, resulting in a reduction of the amount of out-of-focus signal from above and below the focal plane and produces marked increases in both lateral and axial resolution [55–57]. A 2D image of a small partial volume of the specimen centred around the focal plane (optical section) is generated by performing a raster sweep of the specimen at that focal plane, and a 3D reconstruction of a specimen can be generated by stacking 2D optical sections collected in series. Four-dimensional (4D) studies are, of course, also possible [58].

LSCFM is applicable to all fluorescence microscopy applications. For a review of applications in living cells see Terasaki and Dailey [59]. One interesting application of LSCFM is FISH and reflectance *in situ* hybridization (RISH) [60]. These methods are used for the detection of target molecules using a system of coupled antibodies and fluorophores (in FISH) or colloidal particles (in RISH) [61]. *In situ* hybridization is a mixed-phase reaction where a labelled nucleic acid probe hybridizes to a nucleic acid target. The technique involves the preparation of short sequences of cDNA complementary to the DNA sequence of interest. These cDNA probes are labelled with fluorescent tags or with colloidal particles and once the probes hybridize to the complementary DNA, the area of DNA can be detected. FISH and RISH combined with confocal laser microscopy offer a high degree of spatial information in location nucleic acid sequences and a higher sensitivity in detection [62] (Figure 24.3).

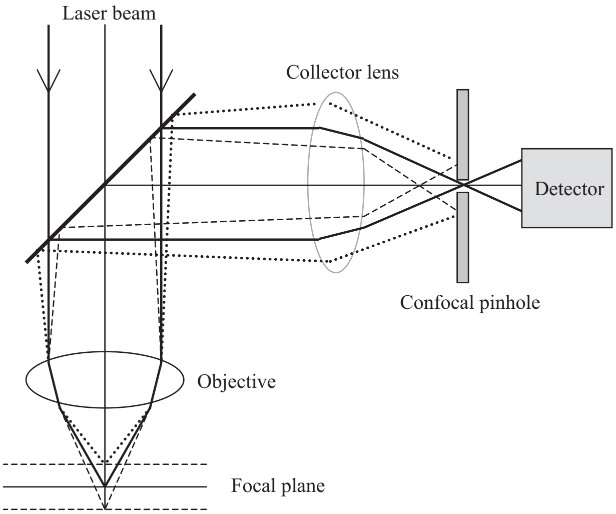


FIGURE 24.3 The geometry of a laser scanning confocal fluorescence microscope.

Even though the resolution of a LSCFM is better than that of conventional microscopes, it is restricted by the diffraction limit of light. According to Abbe's treatment, the dimension d of the smallest object or feature that can be resolved by using a given lens system is

$$d = 0.61\lambda/n\sin\theta = 0.61\lambda/\mathrm{NA} \tag{24.2}$$

where n is the refractive index of the medium between the sample and objective lens; λ is the wavelength of the light; θ is the acceptance angle of the lens; and $n \sin \theta$ is called the *numerical aperture* of the lens, NA. Since the maximum practical value of *NA* for objective lenses operating in air is about 1, this implies that $d_{\min}$, the minimum resolvable feature size is of order

$$\sim 0.61\lambda. \tag{24.3}$$

Consequently, diffraction sets a limit of about half a wavelength to the smallest feature size which can be resolved by conventional optical microscopes and LSCFM techniques. Reducing the wavelength from red light (700 nm) to blue light (350 nm) can improve the resolution only by a factor of two. Using an oil immersion lens ($n > 1.0$) will further improve the resolution by a factor of less than two. In the next section we will discuss other methods aimed to increase the resolution beyond the diffraction limit of light.

24.4.1.2 Laser Microscopy Beyond the Diffraction Limit

a. *Near-field scanning optical microscopy (NSOM).* This method combines scanning probe microscopy instrumentation with optical microscopy and spectroscopy to provide optical characterization with, in some cases, 15 nm resolution using visible light. The technique employs a sharp-end optical fibre that is coated with metal such that a small aperture (approximately 25 nm in diameter) is formed at the tip of the fibre [63–65]. Laser light is directed into the optical fibre and through the aperture to illuminate a small spot on the sample that is much smaller than the conventional diffraction limit. The probe is scanned over a surface at a height above the surface of a few nanometres and fluorescence is collected through the same small aperture. The high resolution that can be achieved with this instrument allows to study, for example, the near membrane fluorescence in order to address questions of near membrane ionic changes. Domains in cell plasma membranes have also been investigated by near-field scanning optical microscopy [66], as well as single DNA molecules [67].

Several books and reviews on applications of NSOM in biology have been published (see, for example, [68–70]).

The contrast in NSOM images can be significantly improved when two-photon excitation is used [71,72]. The excitation is realized by shining laser light on a sharp metal tip. The metal tip provides the local excitation source of the sample under investigation. Excitation light with the appropriate polarization induces a strongly enhanced field at the tip. The enhanced field consists mainly of non-propagating (evanescent) components and is, thus, strongly confined to the tip end. This method may find interesting applications in biology and biotechnology.

b. *Total internal reflection fluorescence (TIRF).* This is based on the total internal reflection resulting when light propagating within a dense medium (e.g. quartz) reaches an interface with a less dense medium (e.g. aqueous solution). Although the light is fully reflected, an evanescent field that extends beyond the interface and into the aqueous solution is generated. Typically, the penetration depth (or the thickness of evanescent field) is in the range of half the wavelength of the light. The evanescent field provides the surface selectivity of TIRF. Only fluorophores adsorbed, adhered or bound to the surface will be excited and, therefore, fluoresce. Conversely, fluorophores in bulk solution will not be excited. Therefore, if the surface is made biologically active so that fluorescently-labelled compounds of interest can be bound to the surface, one can detect analytes within complex sample solutions. Because the excitation light is totally reflected away from the detection, it is easy to discriminate the fluorescence signal from the excitation light and achieve very low sensitivities and detection limits down to the single-molecule level. TIRF systems provide measurement of real-time kinetics of a bioanalyte's binding to a surface immobilized sensor molecule. TIRF has been successfully applied to numerous studies associated with solute adsorption, orientation and rotational mobility associated with conformational changes. Specific applications can be found in Oheim et al. [73], Axelrod et al. [74], Schmoranzer et al. [75], Zenisek et al. [76], Ishijima et al. [77] and Kitamura et al. [78].

c. *Point-spread-function engineering.* Far-field microscopy, including confocal and multi-photon excitation microscopy, can achieve resolutions of about 250 nm in the *x–y* (transverse) and 500 nm in the *z* (axial) directions. Point-spread-function (PSF) engineering [79] improves this resolution fundamentally by reducing the effective PSF of these microscopes in one or several directions. This is accomplished by combining suitable (photo) physical processes with spatial probability patterns of excitation and detection.

In *standing-wave (SW) microscopy* [80], coherent (laser) light interference is used for excitation in a fluorescence microscope in order to enhance axial resolution beyond what is attainable through either computational fluorescence optical sectioning microscopy or confocal/multi-photon scanning fluorescence microscopy. The interference of two beams split from a laser sets up a periodic SW pattern of transverse anti-node planes that span the thickness of the specimen. By control of the position of the SW field planes relative to the specimen, the axial structure that would otherwise be unresolved within the depth-of-field can be separated into distinct image components. Provided that the specimen is only very weakly refractive, the field intensity is a simple function of the axial coordinate (*z*) and uniform over any transverse plane. Because this field excites fluorescence in the specimen proportionally, it can be thought of as a weighting factor of the object, provided the 'phase' of the field is known between focal planes. It can be readily demonstrated that any SW image is a linear combination of three independent images, here taken to be the conventional image, and images of the cosine and sine-weighted objects at field spatial frequency k. Therefore, for a given SW field node spacing (fixed k), three distinct SW images capture all of the available information at each focal plane. In very thin specimens where there are structures such as cytoskeletal fibres, axial shifting of the SW field can be used to trace the contour with a precision as fine as 36 nm. More generally, however, three SW images must be acquired at each focus plane and the three resulting image stacks computationally processed to extract the extra information.

In the concept of *stimulated emission depletion microscopy* [81], Abbe's diffraction resolution barrier is overcome by hindering excited molecules at the outer region of the focal spot from emitting a fluorescence photon. Excitation is performed by a sub-picosecond laser pulse that is followed by a picosecond red-shifted pulse inducing stimulated emission in the excited dye. The net effect of stimulated emission is that the affected molecules are quenched to the ground state immediately after the excitation. By arranging the stimulating light beam in a doughnut mode around the focal maximum, one can inhibit fluorescence just at the periphery of the spot. By driving the quenching process into saturation, one can reduce the spatial region from which fluorescence occurs and, hence, fundamentally break the diffraction barrier [82,83].

24.4.2 Applications Based on Non-linear Optical Responses

24.4.2.1 Multi-photon Laser Scanning Fluorescence Microscopy

Many potentially interesting biological imaging experiments are restricted by the damaging effects of the illumination used, i.e. photobleaching of the probe and toxicity of the photoproducts. In LSCFM (see Section 24.4.1.1), the light used for excitation is in the visible region of the electromagnetic spectrum and is directly absorbed by the biological molecules situated in and out of the focal point of the focusing lens, contributing to rapid photobleaching and photoproduct formation.

The compromise that needs to be made is to reduce the number of sections in a 3D section series and to reduce the average number of scans per image or the number of time points when observing an on-going process. Another disadvantage in one-photon microscopy is the small depth of penetration of visible light in the biological material.

Multi-photon fluorescence microscopy is a powerful method that combines the advantages of laser scanning microscopy with long wavelength multi-photon excitation to capture high-resolution, 3D images of specimens.

Multi-photon excitation, which is based on the simultaneous absorption of photons, was predicted in 1931 [84] and first realized with the availability of lasers in 1961 [85]. Two-photon laser-scanning microscopy (TPLSM) was first introduced by Denk *et al.* in 1990 for studies of living cells [86].

In order to produce excitation, two photons must interact with a fluorophore simultaneously and must have a combined (summed) energy equal to the absorption resonance energy. In other words, the two photons, which are usually in the infrared, will be absorbed by the fluorophore to a state virtually identical to that caused by a single visible photon of half the wavelength. Fluorescence will be emitted as a result of radiative relaxation to the ground state. Because the two-photon process is a nonlinear process, the absorption and fluorescence depend on the square of the infrared light intensity. This nonlinear effect, combined with the sharp focusing of a microscope objective lens, allows excitation of a diffraction-limited spot within the specimen without the background fluorescence produced in the out-of-focus region in front of and behind the beam waist as shown in Figure 24.4.

The intensity of the two-photon excited fluorescence, I_F, follows the relation [86]:

$$I_F \approx \frac{P^2\alpha}{\tau f^2}\pi^2\frac{NA^4}{\left(hc\lambda\right)^2} \tag{24.4}$$

where P is the mean power, α is the molecular two-photon absorption coefficient, τ is the pulse width, f is the repetition frequency, NA is the numerical aperture, h is Planck's constant, c is the speed of light and λ is the wavelength. Because the fluorescence yield depends on a P^2/τ relation, two-photon microscopy with 1 ps laser pulses requires three times higher average power than 100 fs pulses.

A number of studies have established the applicability of TPLSM to fluorescent protein imaging, cultured cells, plants and neurons.

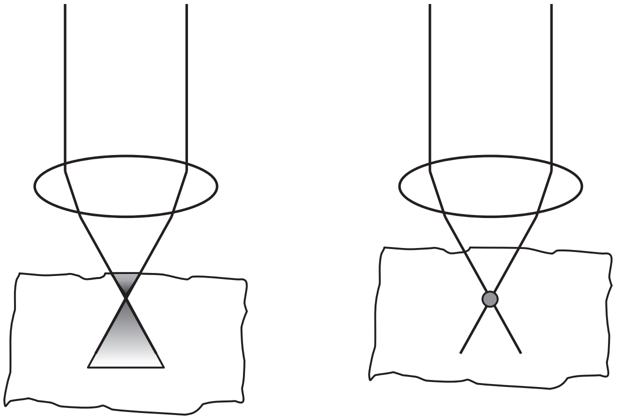


FIGURE 24.4 Comparison of the extent of the focal region giving rise to fluorescence signal (left) in the case of laser scanning confocal fluorescence microscopy (LSCFM) and (right) two-photon laser scanning microscopy (TPLSM).

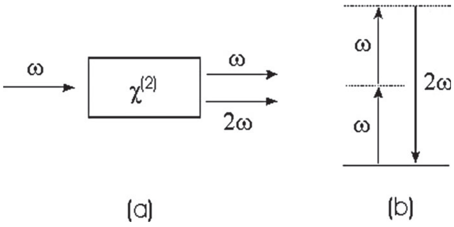


FIGURE 24.5 SHG. (a) If a high-intensity-beam angular frequency ω is applied to a medium with an appreciable second-order susceptibility $\chi^{(2)}$ it can generate radiation at frequency 2ω. The process can be thought of as in (b) the simultaneous absorption of two photons at frequency ω allowing a virtual state of energy 2 $\hbar\omega$ to be accessed momentarily, together with the simultaneous emission of a photon of frequency 2ω.

The localized nature of two-photon absorption lends itself to a number of additional applications. The uncaging process (see Section 24.6.2.2(c)) has been monitored to map acetylcholine-receptor density [87] or cell fates in the developing sea-urchin embryo [88,89]. Others have applied TPLSM to fast calcium imaging and to study the dynamic activity of dendrites. Ultrashort, high-repetition-rate pulsed lasers, like Ti:sapphire lasers, have been lately used in TPLSM. They allow time-resolved fluorescence and pump-probe studies [90]. These and other more recent applications of TPLSM make clear its benefits for problems in biology and biotechnology [91,92].

Three-photon fluorescence microscopy has also been demonstrated [93–96], where three photons are absorbed simultaneously. In this case, I_F depends on a P^3/τ^2 relation, requiring the use of femtosecond laser microscopes. Three-photon excitation has been used to image serotonin in living cells [97] (Figure 24.5).

24.4.2.2 Imaging Techniques Exploiting Second Harmonic Generation (SHG)

SHG has been exploited as a high-resolution non-linear optical imaging method for microscopy of cellular membranes and intact tissues. As the first non-linear microscopic method demonstrated, harmonic imaging has seen a renaissance both in second and third harmonic mode. SHG shares many of the advantageous features for microscopy of another more established non-linear optical technique: two-photon excited fluorescence (see Section 24.4.2.1). Both are capable of optical sectioning to produce 3D images of thick specimens, to penetrate tissue better due to the NIR wavelength used and both result in less photodamage to living tissue than confocal microscopy. SHG is a second-order nonlinear optical process that can only arise from media lacking a centre of symmetry, e.g. an interface such as a membrane, which can, thus, be used to image membrane probes with high specificity. It also displays extraordinary sensitivity in reporting membrane potentials and has the ability to image highly ordered structural proteins without any exogenous labels. SHG is complementary to two-photon excited fluorescence in that it relies on a different contrast mechanism and is most easily detected in the transmitted light optical path. Particularly interesting are efforts to measure changes in cell membrane potential using field-induced SHG in the reflection mode. Intrinsic protein SH signals hold great promise for tissue imaging not requiring contrast agents [98].

24.4.2.3 Imaging Techniques Employing Coherent Anti-stokes Raman Scattering (CARS)

Coherent anti-Stokes Raman scattering (CARS) microscopy is a new method complementary to fluorescence microscopy that allows researchers to localize specific types of molecules inside living cells without artificial dyes or genetic modifications. The CARS method builds on Raman spectroscopy, which chemists have used for decades to create fingerprints for specific molecules.

The first successful CARS imaging experiment took place in 1999 [99]. The coherent amplified anti-Stokes signal provides a more sensitive detection method for specific chemicals inside cells. The researchers have developed and improved the CARS microscopy technique by reducing background noise, while preserving its valuable sensitivity to spectroscopic signatures of specific cellular molecules [100–103].

CARS microscopy has been used to observe apoptosis in action, organelles collapsing into lipid vesicles, nuclei shrinking and cells becoming round [104–108]. Because collagen in fatty tissue surrounding organs changes early in tumour growth, the technique could also be used as an early diagnostic test for certain types of cancer.

24.4.2.4 Surface-Enhanced Raman Scattering

Raman spectroscopy has been applied to biological problems in confocal Raman systems to measure Raman spectra of cellular components with spatial resolution of about 1 μm [109–112]. Raman spectra of DNA in the nucleus of a living cell have also been measured using a laser Raman microscope, where the sample is inside the laser cavity [113]. However, due to high laser excitation power (usually 5–10 mW or more) and the long collection times (hundreds of seconds) required to detect the weak Raman signals, it is difficult to use this method for most biological samples.

SERS is a technique that measures the strongly increased Raman signals from molecules which have been attached to nanometre-sized metallic structures. Although not entirely understood, it is generally agreed that more than one effect contributes to the enhancement of the Raman signal. Two known mechanisms are the electromagnetic and chemical effects [17,114].

The electromagnetic enhancement factor of up to 12 orders of magnitude [115,116] arises from enhanced optical fields due to resonances of the applied optical fields with the surface plasmon oscillations of the metallic nanostructures [117]. Chemical SERS enhancement usually has a very small contribution to the total SERS enhancement and originates from a metal-electron-mediated resonance Raman effect via a charge transfer intermediate state [118]. The desired laser excitation is in the NIR due to the large observed enhancement factors, the lack of fluorescent background and lower risk of photodamage.

A SERS vibrational spectrum may be recorded with shorter integration times, it avoids photodecomposition of the probed molecules because the excitation energy is not in resonance with molecular transitions and at the same time it provides a high degree of structural information about the molecule. The large SERS signal using NIR laser excitation and silver or gold nanocolloids makes it possible to use SERS for single-molecule detection (SMD) as a complementary method to fluorescence [119]. Detecting single molecules in solution with high sensitivity and molecular specificity is of great scientific and practical interest in biology, biotechnology, medicine and, in particular, in human gene analysis, where progress will be greatly enhanced by methods for selective and rapid detection of single molecules [120].

SERS studies have also been performed on living cells to monitor the intra-cellular distribution of drugs in the whole cell and to study the anti-tumour drugs/nucleic acid complexes [121]. Other applications include producing silver nanocolloids within *Escherichia coli* bacteria, resulting in SERS spectra of the constituents forming the cell wall, including proteins, peptides and aminoacids [122].

24.4.3 Spectral-Based Applications—Imaging Beyond the Visible Region

24.4.3.1 Imaging with Soft X-rays in the 'Water-Window' Region

One of the problems encountered when attempting to obtain images of biological samples at much higher resolution than that allowed by the diffraction limit of visible radiation arises from the incompatibility of biological material with the high vacuum requirement of electron microscopes. *In vivo* imaging of cells or tissue by electron microscopy is complicated by the fact that living matter is largely composed of water. Samples must be dried out before they can be put in the electron microscope, which usually leads to a collapse of delicate internal structures. Moreover, it is often necessary to apply metallization or some means of enhancing contrast for the electron scattering process that creates the image, which again complicates interpretation of the image.

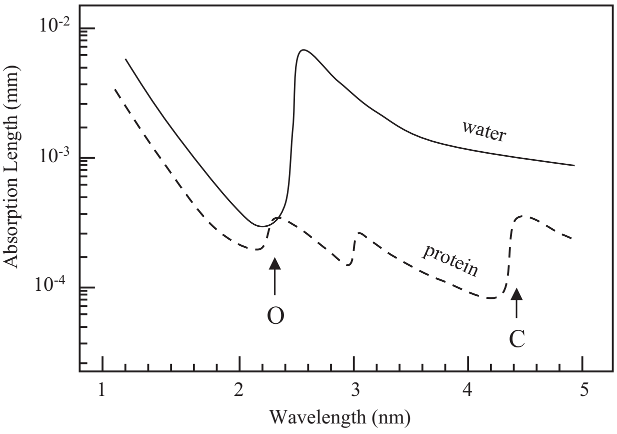


FIGURE 24.6 Schematic diagram of the absorption spectra of water and protein in the soft X-ray region showing the 'water window' between the K-absorption edges of oxygen at 2.34 nm carbon at 4.4 nm.

One possible way of avoiding these problems is to use soft X-rays whose wavelength lies in the so called 'water-window' region of the spectrum bounded by the K-absorption edge of oxygen at 2.34 nm ($h\nu=539$ eV) and the K-absorption edge of carbon at 4.4 nm ($h\nu=282$ eV). For wavelengths in this region, the water (hydrogen and oxygen) is transparent to the X-rays but as shown in Figure 24.6, carbon and heavier elements in proteins and other molecules of living matter are absorbing. The samples can, therefore, be imaged in their *in vivo* wet state. Moreover, there is a natural contrast between the water background and the biological structures within the cells—no staining or metalization of the samples prior to recording the image is necessary.

One technique for imaging biological samples with water-window X-rays exploits the same technology used in the semiconductor industry for fabricating microchips on silicon wafers by photolithography. The wafers are coated with an X-ray-sensitive resist (e.g. PMMA which can be handled in low-level yellow light) and the biological sample in its fully wet state is applied to the resist and covered with a thin window of silicon nitride to separate it from the low vacuum chamber housing the X-ray source. In the case of the system developed in the Clarendon Laboratory, Oxford [123–126], this comprised a laser-produced plasma obtained by focusing a single 2 J shot from a discharge excited KrF laser onto a section of tungsten or molybdenum wire stretched taut in the focal plane of the laser beam. In addition to the filtering effect of the entrance window, a supplementary thin-film filter of vanadium on mylar was used to restrict the spectral range of radiation on the short wavelength since vanadium strongly absorbs radiation of energy higher than 520 eV. The overall effect was that only water-window X-rays from the recombining W or Mo plasma reached the sample.

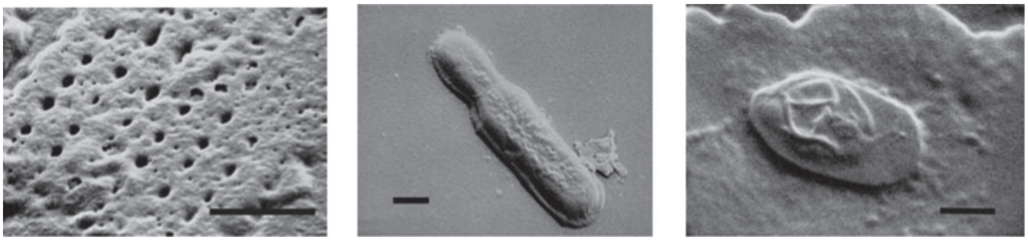
The focal spot on the metal target is typically some 5 μm in diameter. However, the emitting region of plasma is somewhat bigger than this, perhaps 100 μm in diameter since the plasma has expanded somewhat before the peak of the emission pulse is reached. Nevertheless, as it is typically some 10 cm or more distant from the sample, it acts as a quasi-point source producing a sharp contact print on the resist. The resolution is limited only by the penumbral blurring due to finite source size and is comparable to the small degree of blurring resulting from Fresnel diffraction over the sample thickness.

After irradiation by this strong dose of X-rays, delivered in less than 50 ns, the sample will certainly have suffered considerable radiation damage and may be dead. However, the image it will have recorded on the resist in this very brief exposure is that of the sample in its undamaged state, since damage mechanisms cannot affect its appearance on such time scales. The resist on the wafer is developed by conventional techniques such that the areas which have been exposed to a high X-ray flux are dissolved faster than those which have been shielded from the Xx-rays by the organic molecules of the biological specimen. The surface of the developed resist layer bears an image in the form of a contour map of the local water-window X-ray transmission factor of the sample. The silicon wafer is now a dry object that can be scanned in an electron microscope or read out directly (in air) by an atomic force microscope.

Figure 24.7 shows the type of images that can be obtained with this technique. Since the area of silicon wafer which can be exposed on each laser shot is typically 5 mm×5 mm, one exposure is enough to capture 60 000 individual images of a cell of the size shown in Figure 24.7b.

24.4.3.2 Imaging with Lasers Operating in the 'Water-Window' Region

The X-rays used to capture the images in the previous section are essentially incoherent radiation emitted by laser-produced plasmas. There is considerable interest in developing laser sources in the X-ray region especially those which operate



(a) **(b)** **(c)**

FIGURE 24.7 Soft X-ray imaging of biological objects at high resolution. (a) Resolution test of a dry sample—a diatom skeleton. The scale bar is 2 μm long. Features down to about 1/40th of the scale bar, i.e. about 50 nm are resolved [127]. (b) Image of *Schizosaccharomyces pombe in vivo* (fully wet). The scale bar is 2 μm long. Details of the cell wall and septum formation are clearly visible [125]. (c) Image of *Bacillus cereus in vivo*, fully wet specimen. Details of structures within the cell nucleus are clearly resolved. The scale bar is 1 μm long [125].

directly in the 2.3–4.4 nm water window since this would open up new possibilities of holographic imaging and scanning X-ray techniques allowing direct electronic read out of images.

Development of a pulsed soft X-ray source in the water window for investigation of transient inner-shell transitions (e.g. at the carbon K-edge) of biological specimen molecules (XUV/VIS double resonance) has been reported [128].

24.5 Laser-Based Imaging: Applications to Different Levels of Biological Organization

24.5.1 Single-Molecule Detection

SMD probes individual molecules with the potential of revealing fine details about dynamical processes of biological interest in a multitude of *in vitro* systems [129–136]. There are almost no studies *in vivo* involving single molecules, mainly because of the difficulties related with measurement of very small signals in a dynamic system. One report discusses the detection of individual fluorescence-labelled human cardiac L-type Ca^2+ channels, indicating that L-type Ca^{2+} channels tend to form larger aggregates that are mobile in the plasma membrane [137].

In order to be able to detect single molecules, only one molecule of interest should be in the volume under study and this molecule should be distinguishable from the other molecules of the environment. To fulfil these conditions, very low concentrations (on the order of 10^{-10} M) and molecules that interact with light (absorption, emission or SERS) much stronger than the environment are used. Further, the background noise coming from the environment is minimized by focusing laser light in very small volumes. Single molecules can also be studied using optical traps created by focused laser beams.

During the development of SMD, the field has emerged rapidly with many interesting applications including DNA mechanics (twists, turns and stretches) and molecular motors. Ishijima et al. [77] used a combined optical trap–TIRF setup to observe mechanical and ligand-binding events simultaneously in a single myosin molecule. The protein converts chemical energy (adenosine triphosphate ATP hydrolysis to adenosine diphosphate ADP) into kinetic energy with an efficiency of about 50%. Ishijima et al. (1998) [77] used optical traps to detect mechanical binding of myosin to actin and measured simultaneously (with TIRF) the fluorescence from fluorescently labelled ATP. They demonstrated a correlation between the myosin–ATP binding and releasing cycle, and the mechanical cycle of binding and releasing actin. A delay of hundreds of milliseconds between releasing the nucleotide (probably ADP) and force generation has been explained by the existence of a memory state of myosin, which stores chemical energy from ATP hydrolysis after bound ATP is released.

24.5.2 Imaging at Sub-cellular Level

Researchers employed confocal microscopy in combination with exogenously applied fluorescent dyes that preferentially labelled either naked DNA or complexed DNA to discover that DNA may be efficiently transferred through the cell membrane, cytoplasm and nuclear membrane by cationic liposomes. Trafficking of DNA to the nucleus does not depend on disruption of the nuclear membrane during the M phase of cell division as previously proposed. Complexed DNA was found in the cytoplasm and extending through the nuclear membrane but only naked DNA was found in the nucleus, suggesting that the trafficking of DNA from the cytoplasm into the nucleus is facilitated by fusion of the cationic liposome with the nuclear membrane leading to the release of naked DNA into the nucleus. These observations provide a new model for use in developing and optimizing non-viral DNA transgene delivery vehicles which are easy to prepare, possess low immunogenicity and toxicity, may be actively targeted and have an excellent safety profile for gene therapy applications [138].

24.5.3 Imaging of Organelles and Bacteria

SHG has been used to produce very high-resolution images of cellular membranes, voltage-sensitive dyes, centrosomes, tubulin, collagen, muscle and bone in 550 μm thick tissue sections. These images were acquired without photobleaching because this technique relies on non-absorptive optical processes [98].

24.5.4 Imaging at Cellular Level

High-resolution full-field microscopy, considered traditional by now, is still the dominant approach especially in its digital version [54,139], continuing to yield highly useful results; however, laser-based methods, as previously discussed, are gaining significant ground, especially when relying on the competitive advantages in resolution and/or specificity they offer.

Two-photon microscopy was employed to perform rapid imaging of calcium transients in deeply anaesthetized rats. In these experiments, a metal plate was attached by a dental acrylic to the skull of each anaesthetized rat. The skull was immobilized under the microscope objective by two clamps attached to the metal plate. A micro-electrode or patch pipette was inserted through an agar-filled cranial window into the brain at a shallow angle while applying high positive pressure to prevent the tip from clogging and subsequently used at low pressure to fill individual neurones with a calcium indicator and to establish whole cell recording configuration. This technique was combined with electrical tail pinch and whisker stimulation protocols to record calcium transients from a wide variety of neurons in all cortical layers in anaesthetized rats including barrel cortex neurons, pyramidal neurons in somatosensory cortex, and mitral cells and tufted cells in olfactory bulb. The inherent resolution and agility of this technique present the exciting possibility of identifying and measuring changes in dendritic spines representing synaptic plasticity following electrical or somatosensory stimulation [140].

The two-photon imaging technique in fully anaesthetized and restrained rats was extended to freely moving animals using a 25 g fibre microscope that may be attached to the head of adult rats. This tiny microscope enabled visualization of single neurons filled with a calcium indicator, calcium transients in dendrites, small blood vessels and capillary blood flow at depths to 250 μm below the surface of the brain. While technical improvements will be required to increase depth of penetration, this technique can be used to obtain stable images from fully conscious rats engaged in freely moving behavioural tasks. Nevertheless, the ability to visualize action potentials by means of calcium imaging in single neurons in fully intact freely moving animals substantially extends the range of potential neurobiological studies [141].

24.5.5 Imaging in Tissue

Confocal microscopes capable of scanning large specimens (from 0.5 to several centimetres in diameter) were previously implemented using mechanical scanning stages that physically moved the subject to be imaged in a raster scan pattern under a fully stationary vertical optical beam. Such microscopes achieve a lateral resolution of 0.25 μm and a vertical resolution of 0.5 μm. Because there are few optical components in the beam line and the beam remains on axis at all times, light collection efficiency is maximized and aberrations are minimized. The primary limitation on the use of these instruments for imaging large biological specimens is that it usually takes several minutes to acquire each confocal slice.

Standard confocal microscopes are implemented using a beam that is scanned over a stationary subject. These microscopes are capable of achieving the same resolution as scanning stage microscopes. However, due to the use of additional optical components in the beam line and the off-axis movement of the beam, light collection efficiency is reduced and aberrations are increased. The principal advantage *versus* scanning stage microscopes is that scanning beam microscopes can image an entire specimen in a fraction of a second. However, the practical field of view limits subject size to about 0.5–1 cm in diameter.

A new confocal microscope (MacroScope®) was developed incorporating a large *f*–theta telecentric lens that maintains the converging cone of rays perpendicular to the specimen plane and facilitates acquisition of data at equally spaced points across the image. One or more secondary parfocal infinity corrected microscope lenses may be switched into the beam line for acquisition of medium to very high resolution confocal images. This microscope has a 23 cm working distance in macroscopic mode, 5 μm lateral resolution, 200 μm depth resolution and a field of view variable from 25 μm×25 μm to 7.5 mm×7.5 mm [142].

While the original Macroscope® was designed to operate exclusively in reflected light and fluorescence mode, additional capabilities (*n*-photon, lifetime, multi-spectral, optical trapping, laser tissue ablation) may be incorporated into this instrument by the insertion of additional optical elements in the beam line under computer control [40]. These additional capabilities would extend the potential range of applications of this instrument for *in vivo* biological studies.

Nevertheless, the original Macroscope® is capable of producing confocal images at micro-, meso- and macro-scales of intact flies, eyes, leaves, larvae, embryos and transparent fish. While it is possible to obtain similar images of some superficial features using a scanning electron microscope, the specimens would have to be killed, desiccated, coated with a thin conductive coating and viewed in a vacuum. In contrast, the Macroscope® is able to image unprepared specimens, attached to a stage and viewed in air or water. These capabilities make it possible to repeatedly acquire images from living biological specimens in multiple modalities over either short or long periods of time.

Developmental biologists have spent many decades studying the effects of lesions of small groups of cells in embryos on subsequent development. In the past couple of decades, it has become possible to introduce a fluorescent label into cells which is incorporated into their progeny in order to determine the ultimate contribution of progenitor cells to the development of the embryo. For example, by selective lesioning and/or labelling it has become possible to identify specific early embryonic cells giving rise to specific adult tissues or to identify the role of specific cells in the subsequent development of the adult nervous system [143].

Unfortunately, many progenitor cells of interest are buried deep inside embryos and are, therefore, inaccessible to manipulation without causing potentially unnoticed damage to supervening cells. The Macroscope® with an integrated tissue ablation (laser scissor) or optical trapping (laser tweezer) system presents developmental biologists with a nearly perfect tool for highly selective ablation of progenitor cells or the introduction of fluorescent labels at previously inaccessible sites with little or no damage to any supervening tissues. These manipulations may be performed *ex vivo* on pre-implantation embryos which are returned to a host mother for further development in their native environment or performed on post implantation embryos *in vitro*. Subsequent imaging of anatomical changes and physiological effects in the developing embryo using enhanced multi-modal Macroscope® would substantially enlarge the amount of information available and reduce the number of animals required for studies of embryo development.

24.6 Lasers for Micromanipulation of Biological Samples

24.6.1 Opto-Mechanical Manipulation of Biological Samples

24.6.1.1 *Optical Tweezers*

Optical tweezers (OTs) use light to manipulate microscopic objects as small as a single atom. The radiation pressure from a focused laser beam is able to trap small particles. In the biological sciences, these instruments have been used to apply forces in the pN range and to measure displacements in the nm range of objects ranging in size from 10 nm to over 100 μm. As early as 1970, lasers were used to trap various atoms and particles. The first optical traps used visible light. Starting in the mid-80s, optical traps were used also to trap biological particles. Arthur Ashkin is thought to be the pioneer of the OT. His AT&T Bell Laboratories group trapped individual viruses, *E. coli* bacteria and cells as early as 1987 [144,145].

In an optical trap, a laser beam is focused by a high numerical aperture microscope objective to a spot in the specimen plane. The gradient forces generated by the refraction of laser light on the particle hold a small particle at the centre of the beam. Most frequently, OTs are built by modifying a standard optical microscope. These instruments have evolved from simple tools to manipulate micron-sized objects to sophisticated devices under computer control that can measure displacements and forces with high precision and accuracy [146].

OTs have been used to trap dielectric spheres, viruses, bacteria, living cells, organelles, small metal particles and even strands of DNA. Applications include confinement and organization (e.g. for cell sorting), tracking of movement [147] (e.g. of bacteria), application and measurement of small forces and altering larger structures (such as cell membranes). Two of the main uses for optical traps have been the study of molecular motors [148,149] and the physical properties of DNA (Bustamante's group). In both areas, a biological specimen is biochemically attached to a micron-sized glass or polystyrene bead that is then trapped.

OTs were combined with other optical and mechanical techniques [150,151] to study physical properties of DNA and proteins (including unzipping, unfolding and refolding), and receptor–ligand bonding interactions. Mechanical properties such as elasticity, stiffness, rigidity and torque of cells, organelles and single molecules can be measured using OTs [152].

The movement of kinesin along the microtubules has been studied and quantified corresponding to hydrolysis of ATP [153–155]. For other applications see Section 24.5.1.

24.6.1.2 Perforation (Optoporation)

By creating, with laser light, a micron or sub-micron size hole in the membrane of cells (procedure called optoporation), any molecule can be inserted into a cell. The hole usually seals within a fraction of a second, posing no permanent damage to the membrane. Both diode and holmium:yttrium scandium gallium garnet (Ho:YSGG) lasers have been used for the precise drilling of 5–20 μm holes in the zone pellucida of mouse zygotes [156,157]. The procedure was shown to be simple, non-invasive and non-toxic.

Near-infrared ultrashort laser pulses have been used to deliver foreign DNA into cells *in vitro* [158]: 100% transfection efficiency has been reported for different cell types, which sets a landmark in targeted gene therapy and DNA vaccination. A Ti:sapphire femtosecond laser with average power of 50–100 mW and a repetition rate of 80 MHz in the NIR is focused into a sub-femtolitre volume at the cell membrane, creating a tiny, localized perforation in the membrane through which DNA can enter the cell. The integration and expression efficiency of the encoding GFP gene *in situ* was monitored by time-lapse two-photon fluorescence imaging and by two-photon fluorescence-lifetime imaging.

24.6.1.3 Nanosurgery (Laser Scalpels)

A human chromosome was first dissected by a laser in 1986 [159]. Argon-ion lasers were used to micro-dissect plant chromosomes [160]. Later, femtosecond lasers in the NIR were used to perform surgery of cellular nanostructures. They have the advantage of a low ablation threshold and limited thermal damage to surrounding structures. Because only the central part of the Gaussian illumination beam provides sufficient intensity for plasma-induced ablation, laser cuts and holes can be generated in the target with a size below the diffraction-limited spot size. Koenig's group [161,162] realized a minimum cut-size of 110 nm into the human chromosome 1 and showed dissections of chromosomes and knocking out intracellular structures in single living cells at mean powers of 30–50 mW. Intra-tissue nanosurgery of the cell wall in a living leaf of *Elodea densa* has also been successfully accomplished. Interestingly, using the same NIR femtosecond laser pulses, a single plastid or a part of the organelle could be completely knocked out without affecting the adjacent organelles or the viability of the cell.

Such highly precise NIR laser-based nanoprocessing of cellular structures without compromising the vitality of cells has numerous potential applications in cell and developmental biology, particularly in studies addressing spatial-temporal control of developmental events and functional interactions between organelles, as well as cell–cell communication.

24.6.1.4 Tissue Welding and Soldering

Bonding the edges of human tissue is a vital step in most surgical procedures. There has been an increased interest in using lasers in tissue welding (heating the adjacent edges of cuts in tissue by a laser beam) and soldering (applying a biological solder onto the adjacent edges and heating the solder and the underlying tissue) [163,164]. Because these methods are less invasive, the patient experiences reduced tissue trauma and less reaction to foreign bodies, contributing to faster wound healing and better cosmetic results.

The precise control of the temperature of incisions around 60°C seems to be the key point in the successful use of lasers in welding and soldering. Temperature of 40°C or lower causes no effect, while heating incisions to temperatures higher than 70°C may cause thermal damage. Addition of protein solder and chromophores (tissue soldering) seems to increase wound tensile strength while decreasing peripheral tissue damage [65].

CO_2 laser light delivered through an optical fibre heats the tissue, while the temperature is monitored with an IR detector [165,166]. The system is promising for both skin and corneal incisions.

24.6.2 Physico-Chemical Manipulation of Biological Samples

24.6.2.1 Pump-Probe Technique

A frequently used method in time-resolved spectroscopy for investigating transient species is the pump-probe method. This method is used especially to understand fundamental problems like the conversion of light energy into chemical energy or the fundamentals of the light–sample interaction. A pump pulse with a high energy density interacts with the molecule under investigation and excites it to a higher state. The new state is a non-equilibrium one and will tend to achieve the potential energy minimum by relaxation. A second probe pulse of low energy density, tuned to a wavelength so that it will be able to monitor the molecule in the new state, is spatially overlapped with the pump pulse with an adjustable delay time. By changing the delay between the pump and probe, one can monitor the changes taking place in the molecule. The characteristic spectra obtained from these transient species give specific information about the dynamics and changes in the molecular structure during the transition. The delay between the two pulses can be easily changed by varying the optical path for one of the two pulses with the help of delay lines which can achieve 1 μm resolution. For example, in air the light waves need around 3.3 ps to propagate through 1 mm and about 3.3 fs through 1 μm.

The excitation of the sample molecules with ultra-short (femtosecond) pump pulses creates coherent vibrational excitation of the Franck–Condon-active modes that have a period of vibration comparable or longer than the duration of the laser pulse. The delayed probing pulse monitors the subsequent dynamics of these vibrational modes. If the pump laser is resonant to an optically allowed electronic transition in the molecule, coherence is generated in both ground and excited state vibrational manifolds. Experimentally, the coherent effects are seen as oscillations in time of the measured spectroscopic property (e.g. transient absorption). These oscillations decay, in solutions, on a fs to ps time scale, due to anharmonic coupling to other vibrational modes.

24.6.2.2 Laser-Activated Processes

a. *Fluorescence recovery after photobleaching (FRAP).* FRAP measurements are used to quantify a number of dynamic parameters of fluorescent molecules, including diffusion coefficient, immobile fraction and binding or residence time within a medium [167]. A region with fluorescent molecules is irradiated and photo-bleached with laser light. The subsequent redistribution (recovery) of fluorescent and bleached molecules throughout the volume of the sample is monitored, giving information about the mobility of the fluorophores. These fluorophores may be exogenous fluorophores introduced into a system specifically for the purpose of the FRAP measurement, such as to measure membrane fluidity or maybe endogenous fluorescent molecules which are themselves of interest. FRAP can be used in combination with one- or multi-photon laser excitation (see multi-photon processes).
b. *Chromophore-assisted laser inactivation (CALI).* This is a light-mediated technique used to selectively inactivate proteins within cells. This method uses laser light to inactivate specific proteins within a cell to learn more about the functions of these proteins. Malachite green is conjugated with non-blocking antibodies that further specifically bind to proteins. When excited with laser light, the dye generates short-lived free radicals capable of damaging proteins.
c. *Uncaging.* Uncaging is an important tool in cell biology and biotechnology, providing the means for precise control of compound delivery with fine temporal and spatial resolution while observing the consequences for cellular processes. Intense UV light from lasers or flashlamps is used in these experiments.

A wide range of bioactive molecules can now be conjugated with caging groups, including transmitters such as glutamate, messengers such as IP3, intra-cellular cAMP and calcium [168–172] as well as other molecules including nucleotides [173]. The caged bioactive molecule is inert until the cage opens by photolysis. The time and position in space of photolysis can be controlled by the application of localized intense UV light generated by lasers or flashlamps. The applications include studies of the operation of receptors [174,175], visual transduction [176], gene expression [177], synaptic transmission [168], astrocyte-neuron signalling [178] as well as cardiac [179] and neuronal [180,181] calcium channels.

Multi-photon excitation has been successfully used for uncaging [182–184]. The unique feature of multi-photon uncaging is that it can provide true 3D localization of excitation in volumes down to 1 femtolitre within cells or tissue. The limitation of this technique is the low cross section of the two-photon process for the commonly used caging groups.

24.6.3 Fluorescence Correlation Spectroscopy

Fluorescence Correlation Spectroscopy (FCS) measures spontaneous fluorescence intensity fluctuations with time as molecules move in and out of the excited microscopic volume of about 10^{-15} l (1 fl) defined by a tightly focused laser beam. The interest in FCS has been stimulated by the fact that it is inherently miniaturized and, therefore, applicable for high-throughput screening applications. Small, rapidly diffusing molecules produce rapidly fluctuating intensity patterns, whereas larger molecules produce more sustained bursts of fluorescence. The time pattern (autocorrelation) can give information about the rate of diffusion of molecules and their concentration in the analysis volume. Visible wavelength cw laser sources can be used as excitation light. Those include the 488 nm argon-ion spectral line, the 543 and 633 nm HeNe laser spectral lines and the 568 and 647 nm ArKr laser spectral lines. Diffusion processes, drug binding, protein folding and singlet/triplet kinetics can be studied. Specific applications include nucleic acid fragmentation [185–187] and hybridization [188], lateral segregation of lipids in bilayer membranes [189], molecular diffusion in the nucleus and cytoplasm [190], protein-protein interactions [191–194], and clustering of membrane–bound receptors [195].

24.6.4 Flow Cytometry

Flow cytometry (FC) involves the use of a beam of laser light projected through a liquid stream that contains cells or other particles. In order to make the measurement of biological/biochemical properties of interest easier, the cells are usually stained with fluorescent dyes that bind specifically to cellular constituents. The cells/particles are confined to the centre of a laminar flow stream that is intersected by the illuminating beam at the interaction region. Those particles that pass through the focused laser beam emit fluorescence or scatter light. The light originating from the particles as they pass through the interaction region is collected in the forward and perpendicular directions relative to the illumination beam by specifically designed optics. This signal is then converted for computer storage and data analysis and can provide information about various cellular properties. FC can measure physical characteristics such as cell size, shape and internal complexity. Moreover, any cell component or function that can be detected by a fluorescent compound can be examined. This technology has been reviewed by many, none better than Shapiro [196].

Applications include (i) cell cycle analysis by measuring the DNA content of the cells; (ii) cell apoptosis by measuring the changes in plasma membrane, the analysis of cell organelles or the sensitivity of DNA to denaturation; (iii) multi-colour analysis for detecting if antigens co-exist on the same cells; (iv) cell sorting based on antigen expression, GFP expression, DNA content or cell function and (v) functional analysis, as determined by measurements of the intracellular calcium level (the dye usually used is UV-excited Indo-1), enzyme activity (by loading cells with a non-fluorescent substrate which may be switched on by an enzyme to yield a fluorescent product) or cell proliferation (by measuring the mitosis-induced fluorescence reduction of a lipophilic dye integrated into the cell membrane).

An interesting new development in FC is multi-colour analysis [197,198], where (a tour de force) up to 11 colours have been detected [199]. This technique provides a large amount of novel information for each sample studied, especially for rare cell populations, and allows identification and characterization of novel cell subsets. Moreover, it may help to identify functionally homogeneous subsets of cells within the complex immune system which is important for understanding disease pathogenesis.

24.7 Emerging Applications of Lasers in Biotechnology

24.7.1 DNA Micro-arrays and Genomics

As a result of the Human Genome Project, there is a large amount of information available about the DNA sequence of the human genome.

The new challenge that researchers face now is the identification and classification of this DNA sequence information and the assignment of functions to these new genes. This is where the DNA micro-array technology (or gene chips) can help. A DNA micro-array works by exploiting the ability of a given mRNA molecule to bind specifically to, or hybridize to, the DNA template from which it originated. By using an array containing many DNA samples, scientists can determine, in a single experiment, the expression levels of hundreds or thousands of genes within a cell by measuring the amount of mRNA bound to each site on the array. With the aid of a computer, the amount of mRNA bound to the spots on the micro-array is precisely measured, generating a profile of gene expression in the cell.

An array is an orderly arrangement of samples. The sample spot sizes in a micro-array are typically less than 200 μm in diameter and these arrays usually contain thousands of spots. The spots are fabricated by high-speed robotics, generally on glass or polymer substrates, and consist of probes with known identity. These probes are used to determine complementary binding, thus allowing parallel gene expression and gene identification studies [200].

In a DNA micro-array experiment, the DNA chip is incubated with the hybridization mixture containing fluorescently labelled cDNAs. The fluorescence of the bound cDNA may be detected using LSCFM. The lasers frequently used for fluorescence excitation are the argon-ion and HeNe lasers. They may allow to more than 50 dyes to be excited and detected at the same time. The results are analyzed using computational methods. Laser stability is the big requirement in these experiments.

Applications are in gene discovery, disease diagnosis, drug discovery (Pharmacogenomics), toxicological research (Toxicogenomics) [201–205].

24.7.2 Protein Micro-arrays and Proteomics

Proteomic research focuses on two main problems: (i) quantification of all the proteins expressed in a cell and (ii) functional study of thousands of proteins in parallel. For quantification purposes, the standard method is 2D gel separation followed by mass spectrometry identification. For protein function studies, micro-array-based assays are being used to study protein–protein and protein–ligand interactions.

The idea of a protein micro-array is not new. The basics and theory of protein micro-arrays were developed in the 1980s by Roger Ekins and colleagues [206,207]. The technique used to detect fluorescence from protein micro-arrays is the same as in DNA micro-arrays. Some applications can be found in [208–212].

24.7.3 Bio-Cavity Lasers

A very promising application of lasers in biomedicine offers the bio-cavity laser. It employs a novel method, developed by Gourley et al. [213–216] at Sandia National Laboratories, to analyze various substances such as blood, cells or tissue. It uses a semiconductor wafer in which the sample acts as an internal component and produces lasing. Although it is simple in design, the profound number of applications that can be derived from this instrument is considerable.

The bio-cavity laser is based on vertical-cavity surface-emitting laser technology. The biological sample is placed between the cavity mirrors inside the resonator of a semiconductor laser. The semiconductor is photopumped by a separate laser to generate electron–hole pairs which recombine to emit light. The unique design of this device allows living cells (red or white blood cells) or other particles to act as transparent waveguides to channel the light between mirrors and become an active part in the lasing process. The cells alter the spectrum of emitted light, based on their absorption and change in refractive index. These variations provide spectral information that distinguishes, for example, malignant or anaemic cells from normal. The bio-cavity laser can function as a kind of lab-on-a-chip, bringing fluids into a micro-laboratory and reading results on the spot. A sickle-shaped red blood cell is clearly distinguishable from normal, spherical cells. The components of a white blood cell, held inside the cavity, can be analyzed to see how it reacts to organisms or drugs in the blood, thus aiding in the design of new drugs. The sample may be whole blood, cells from culture, tissue from a biopsy or a scraping of body tissue.

The emerging applications, together with the more established ones, make it clear that the future of lasers in biology and biotechnology is very bright indeed. What is mostly needed is a true collaboration between those developing relevant laser technologies and those who can apply these to reaching important research goals, in a strategic, interdisciplinary setting.

REFERENCES

1. Shonat R. D., Wachman E. S., Niu W., Koretsky A. P. and Farkas D. L. 1997 Near-simultaneous hemoglobin saturation and oxygen tension maps in mouse brain using an AOTF microscope *Biophys. J.* **73** 1223–31.
2. Shonat R. D., Wachman E. S., Niu W., Koretsky A. P. and Farkas D. L. 1998 Near-simultaneous hemoglobin saturation and oxygen tension maps in mouse cortex during amphetamine stimulation *Adv. Exp. Biol. Med.* **454** 149–58.
3. Breton J. and Geacintov N. E. 1980 Picosecond fluorescence kinetics and fast energy transfer processes in photosynthetic membranes *Biochim. Biophys. Acta* **594** 1–32.
4. Holzwarth A. R. 1989 Applications of ultrafast laser spectroscopy for the study of biological systems *Q. Rev. Biophys.* **22** 239–326.
5. Papagiannakis E., Kennis J. T., van Stokkum I. H., Cogdell R. J. and van Grondelle R. 2002 An alternative carotenoid-to-bacteriochlorophyll energy transfer pathway in photosynthetic light harvesting *Proc. Natl Acad. Sci. USA* **99** 6017–22.
6. Yao B., Xu D., Hou X., Hu K. and Wang A. 2003 Kinetics of picosecond laser pulse induced charge separation and proton transfer in bacteriorhodopsin *J. Biomed. Opt.* **8** 48–52.
7. Lakowicz J. R. 1999 *Principles of Fluorescence Spectroscopy* (New York: Plenum).
8. Marcu L., Cohen D., Maarek J. M. and Grundfest W. S. 2000 Characterization of type I, II, III, IV, and V collagens by time-resolved laser-induced fluorescence spectroscopy *Optical Biopsy III, SPIE Proc.* **3917** 93–101.
9. Marcu L., Fishbein M., Maarek J. M. and Grundfest W. S. 2001 Discrimination of human coronary artery atherosclerotic lipid-rich lesions by time-resolved laser-induced fluorescence spectroscopy *Arterioscler Thromb. Vasc. Biol.* **21** 1244–50.
10. Mayinger B., Jordan M., Horner P., Gerlach C., Muehldorfer S., Bittorf B. R., Matzel K. E., Hohenberger W., Hahn E. G. and Guenther K. 2003 Endoscopic light-induced autofluorescence spectroscopy for the diagnosis of colorectal cancer and adenoma *J. Photochem. Photobiol. B* **70** 13–20.
11. Millot C., Bondza-Kibangou P., Millot J. M., Lallemand A. and Manfait M. 2003 Autofluorescence spectroscopy of malpighian epithelial cells, as a new tool for analysis of cervical cancer precursors *Histol. Histopathol.* **18** 479–85.
12. Shibuki K., Hishida R., Murakami H., Kudoh M., Kawaguchi T., Watanabe M., Watanabe S., Kouuchi T. and Tanaka R. 2003 Dynamic imaging of somatosensory cortical activity in the rat visualized by flavoprotein autofluorescence *J. Physiol.* **549** 919–27.
13. Heim R. and Tsien R. Y. 1996 Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer *Curr. Biol.* **6** 178–82.
14. Miyawaki A., Llopis J., Heim R., Mccaffery J. M., Adams J. A., Ikura M. and Tsien R. Y. 1997 Fluorescent indicators for Ca^{2+} based on green fluorescent proteins and calmodulin *Nature* **388** 882–7.
15. Zhang J., Campbell R. E., Ting A. Y. and Tsien R. Y. 2002 Creating new fluorescent probes for cell biology *Nature* **3** 906.
16. Bruchez M., Moronne M., Gin P., Weiss S. and Alivisatos A. P. 1998 Semiconductor nanocrystals as fluorescent biological labels *Science* **281** 2013–16.
17. Moskovits M. 1985 Surface-enhanced spectroscopy *Rev. Mod. Phys.* **57** 783–826.
18. Jeanmaire D. L. and Van Duyne R. P. 1977 Surface Raman spectroelectrochemistry Part I. Heterocyclic, aromatic, and aliphatic amines adsorbed on the anodized silver electrode *J. Electroanal. Chem.* **84** 1–20.
19. Creighton J. A., Blatchford C. G. and Albrecht M. G. 1979 Plasma resonance enhancement of Raman scattering by pyridine adsorbed on silver or gold sol particles of size comparable to the excitation wavelength *J. Chem. Soc. Faraday Trans. 2: Mol. Chem. Phys.* **75** 790–8.
20. Weitz D. A., Garoff S., Gersten J. I. and Nitzan A. 1983 The enhancement of Raman scattering, resonance Raman scattering and fluorescence from molecules adsorbed on a rough silver surface *J. Chem. Phys.* **78** 5324.
21. Felidj N., Aubard J., Lévi G., Krenn J. R., Salerno M., Schider G., Lamprecht B., Leitner A. and Aussenegg F. R. 2002 Controlling the optical response of regular arrays of gold particles for surface-enhanced Raman scattering *Phys. Rev. B* **65** 075419.
22. Felidj N., Aubard J., Levi G., Krenn J. R., Schider G., Leitner A. and Aussenegg A. R. 2000 Enhanced substrate induced coupling in two-dimensional gold nanoparticle arrays *Phys. Rev. B* **66** 245407.
23. Gunnarsson L., Bjerneld E. J., Xu H., Petronis S., Kasemo B. and Kaell M. 2001 Interparticle coupling effects in nanofabricated substrates for surface-enhanced Raman scattering *Appl. Phys. Lett.* **78** 802–4.
24. Lakowicz J. R., Szmacinski H., Nowaczyk K., Berndt K. W. and Johnson M. 1992 Fluorescence lifetime imaging *Anal. Biochem.* **202** 316–30.
25. Lakowicz, J. R., Szmacinski H., Nowaczyk K. and Johnson M. L. 1992 Fluorescence lifetime imaging of free and protein-bound NADH *Proc. Natl Acad. Sci. USA* **89** 1271–5.
26. Wagnieres G. A. et al. 1998 Clinical evaluation of the cutaneous phototoxicity of 5,10,15,20-tetra(m-hydroxyphenyl) chlorin *Photochem. Photobiol.* **68** 382–7.
27. Bigio I. J. and Mourant J. R. 1997 Ultraviolet and visible spectroscopies for tissue diagnostics—fluorescence spectroscopy and elastic-scattering spectroscopy *Phys. Med. Biol.* **42** 803–14.
28. Andersson-Engels S., Klinteberg C., Svanberg K. and Svanberg S. 1997 In-vivo fluorescence imaging for tissue diagnostics *Phys. Med. Biol.* **42** 815–24.

29. Richards-Kortum R. and Sevick-Muraca E. 1996 Quantitative optical spectroscopy for tissue diagnostic *Annu. Rev. Phys. Chem.* **47** 555–606.
30. Birch D. J. S. 2001 Multiphoton excited fluorescence spectroscopy of biomolecular systems *Spectrochim. Acta A* **57** 2313–36.
31. Levenson R. M., Balestreire E. M. and Farkas D. L. 1999 Spectral imaging: Prospects for pathology. *Applications of Optical Engineering to the Study of Cellular Pathology* ed. E. Kohen (Trivandrum: Research Signpost) pp. 133–14.
32. Levenson R. M. and Farkas D. L. 1997 Digital spectral imaging for histopathology and cytopathology *Proc. SPIE* **2983** 123–35.
33. Farkas D. L. and Becker D. 2001 Applications of spectral imaging: detection and analysis of human melanoma and its precursors *Pigment Cell Res.* **14** 2–8.
34. Moreton R. B. 1994 Optical methods for imaging ionic activities *Scan. Microsc. Suppl.* **8** 371–90.
35. Taylor C. R. and Cote R. J. 1997 Immunohistochemical markers of prognostic value in surgical pathology *Histol. Histopathol.* **12** 1039–55.
36. Taylor D. L., Burton K., DeBiasio R. L., Giuliano K. A., Gough A. H., Leonardo T., Pollock J. A. and Farkas D. L. 1997 Automated light microscopy for the study of the brain: Cellular and molecular dynamics, development and tumorigenesis *Ann. New York Acad. Sci.* **820** 208–28.
37. Brelje T. C., Wessendorf M. W. and Sorenson R. L. 1993 Multicolor laser scanning confocal immunofluorescence microscopy: Practical application and limitations. *Cell Biological Applications of Confocal Microscopy (Methods in Cell Biology 38)* (San Diego, CA: Academic) pp. 97–181.
38. Herrington C. S. 1998 Demystified *in situ* hybridisation *Mol. Pathol.* **51** 8–13.
39. Speicher M. R., Ballard S. G. and Ward D. C. 1996 Karyotyping human chromosomes by combinatorial multifluor FISH *Nature Genet.* **12** 368–75.
40. Farkas D. L., Baxter G., DeBiasio R. L., Gough A., Nederlof M. A., Pane D., Pane J., Patek D. R., Ryan K. W. and Taylor D. L. 1993 Multimode light microscopy and the dynamics of molecules *Ann. Rev. Physiol.* **55** 785–817.
41. Taylor D. L., DeBiasio R., LaRocca G., Pane D., Post P., Kolega J., Giuliano K., Burton K., Gough A., Dow A., Yu H., Waggoner A. and Farkas D. L. 1994 Potential of machine-vision light microscopy in toxicologic pathology *Toxicol. Pathol.* **22** 145–59.
42. Descour M. R., Volin C. E., Dereniak E. L., Thome K. J., Schumacher A. B., Wilson D. W. and Maker P. D. 1997 Demonstration of a high-speed nonscanning imaging spectrometer *Opt. Lett.* **22** 1271–3.
43. Garini Y., Katzir N., Cabib D., Buckwald R. A., Soenksen D. G. and Malik Z. 1996 Spectral bio-imaging. *Fluorescence Imaging Spectroscopy and Microscopy* (Chemical Analysis Series, vol. 137) ed. X. F. Wang and B. Herman (New York: Wiley).
44. Farkas D. L., Levenson R. M., Ballou B. T., Du C., Lau C., Wachman E. S. and Fisher G. W. 1998 Non-invasive image acquisition and advanced processing in optical bioimaging *Comput. Med. Imag. Graph.* **22** 89–102.
45. Wachman E. S., Farkas D. L. and Niu W. 1998 Light microscope having acousto-optic tunable filters, US Patent 5,841,577.
46. Wachman E. S., Farkas D. L. and Niu W. 1998 Submicron imaging system having an acousto-optic tunable filter, US Patent 5,796,512.
47. Wachman E. S., Niu W. H. and Farkas D. L. 1996 Imaging acousto-optic tunable filter with 0.35-micrometre spatial resolution. *Appl. Opt.* **35** 5520–6.
48. Farkas D. L., Ballou B. T., Fisher G. W. and Fishman D. 1996 Microscopic and mesoscopic spectral bio-imaging *Proc. SPIE* **2678** 200–9.
49. Farkas D. L., Ballou B. T., Fisher G. W. and Taylor D. L. 1995 From in vitro to in vivo by dynamic multiwavelength imaging *Proc. SPIE* **2386** 138–49.
50. Galbraith W., Wagner M. C., Chao J., Abaza M., Ernst L. A., Nederlof M. A., Hartsock R. J., Taylor D. L. and Waggoner A. S. 1991 Imaging cytometry by multiparameter fluorescence *Cytometry* **12** 579–96.
51. Kirkwood J. M., Farkas D. L., Chakraborty A., Dyer K. F., Tweardy D. J., Abernethy J. L., Edington H. D., Donnelly S. S. and Becker D. 1999 Systemic interferon-treatment leads to Stat3 inactivation in melanoma precursor lesions *Molec. Med.* **5** 11–20.
52. Schrueck E., du Manoir S., Veldman T., Schoell B., Wienberg J., Ferguson-Smith M. A., Ning Y., Ledbetter D. H., Bar-Am I., Soenksen D., Garini Y. and Ried T. 1996 Multicolor spectral karyotyping of human chromosomes *Science* **273** 494–7.
53. Yang P., Farkas D. L., Kirkwood J. M., Abernethy J. L., Edington H. D. and Becker D. 1999 Macroscopic spectral imaging and gene expression analysis of the early stages of melanoma *Molec. Med.* **5** 785–94.
54. Farkas D. L., Ballou B., Fisher G. W., Lau C., Niu W., Wachman E. S. and Levenson R. M. 1997 Optical image acquisition, analysis and processing for biomedical applications Springer Lecture Notes in Computer Science, vol. 1311, pp. 663–71.
55. Sheppard C. J. R. 1989 Axial resolution of confocal fluorescence microscopy *J. Microsc.* **154** 237–1.
56. Wilson T. 1986 Confocal light microscopy *Ann. New York Acad. Sci.* **483** 416–27.
57. Wilson T. and Sheppard C. 1984 *Theory and Practice of Scanning Optical Microscopy* (London: Academic Press).
58. Thomas C., Devries P., Hardin J. and White J. 1996 4-dimensional imaging—computer visualization of 3D movements in living specimens *Science* **273** 603–7.
59. Terasaki M. and Dailey M. E. 1995 Confocal microscopy of living cells. *Handbook of Confocal Microscopy* 2nd edn, ed. J. Pawley (New York: Plenum) pp. 327–46.
60. Wilkinson D. G. 1994 *In Situ Hybridization. A Practical Approach* (Oxford: IRL Press).
61. Linares-Cruz G., Rigaut J. P., Vassy J., de Oliveira T. C., de Cremoux P., Olofsson B. and Calvo F. 1994 Reflectance *in situ* hybridization (RISH): detection, by confocal reflectance laser microscopy, of gold-labelled riboprobes in breast cancer cell lines and histological specimens *J. Microsc.* **173** 27–38.
62. Makowski E. R. and Ruzin S. E. 1994 Fluorescent in situ hybridization of maize meiotic chromosomes as visualized by confocal microscopy *Biotechniques* **16** 256–8, 260–3.
63. Betzig E. and Trautman J. K. 1992 Near field scanning optical microscopy *Science* **257** 189.

64. Betzig E., Trautman J. K., Harris T. D., Weiner J. S. and Kostelak R. L. 1991 Breaking the diffraction barrier: Optical microscopy on a nanometric scale *Science* **251** 1468.
65. Strassmann E., Loya N., Gaton D., Ravid A., Kariv N., Weinberger D. and Katzir A. 2001 Laser soldering of the cornea in a rabbit model using a control-temperature CO_2 laser system *Proc. SPIE* **4244** 253–65.
66. Hwang J., Gheber L., Margolis L. and Edidin M. 1998 Domains in cell plasma membranes investigated by near-field scanning optical microscopy *Biophys. J.* **74** 2184–90.
67. Garcia-Parajo M., Veerman J. A., Ruiter A. and van Hulst N. 1998 Near-field optical and shear-force microscopy of single fluorophores and DNA molecules *Ultramicroscopy* **71** 331–19.
68. Lewis A., Radko A., Ben-Ami N., Palanker D. and Lieberman K. 1999 Near-field scanning optical microscopy in cell biology *Trends Cell Biol.* **9** 70–2.
69. Pohl D. W. 1990 *Scanning Near-Field Optical Microscopy* (Advances in Optical and Electron Microscopy, vol. 12) ed. C. J. R. Sheppard and T. Mulvey (London: Academic).
70. Zhu X. and Ohtsu M. 1999 *Near-Field Optics: Principles and Applications* (Singapore: World Scientific).
71. Novotny L., Bian R. X. and Xie X. S. 1997 Theory of nanometric optical tweezers *Phys. Rev. Lett.* **79** 645.
72. Sanchez E. J., Novotny L. and Xie X. S. 1999 Near-field fluorescence microscopy based on two-photon excitation with metal tips *Phys. Rev. Lett.* **82** 4014–17.
73. Oheim M., Loerke D., Chow R. and Stuhmer W. 1999 Evanescent-wave microscopy: A new tool to gain insight into the control of transmitter release *Phil. Trans. R. Soc. B* **354** 307–18.
74. Axelrod D. 1999 Surface fluorescence microscopy with evanescent illumination. *Light Microscopy in Biology* ed. A. Lacey (New York: Oxford University Press) pp. 399–423.
75. Schmoranzer J., Goulian M., Axelrod D. and Simon S. 2000 Imaging constitutive exocytosis with total internal reflection fluorescence microscopy *J. Cell Biol.* **149** 23–31.
76. Zenisek D., Steyer J. and Almers W. 2000 Transport, capture and exocytosis of single synaptic vesicles at active zones *Nature* **406** 849–54.
77. Ishijima A., Kojima H., Funatsu T., Tokunaga M., Higuchi H., Tanaka H. and Yanagida T. 1998 Simultaneous observation of individual ATPase and mechanical events by a single myosin molecule during interaction with actin *Cell* **92** 161–71.
78. Kitamura K., Tokunaga M., Iwane A. H. and Yanagida T. 1999 A single myosin head moves along an actin filament with regular steps of 5.3 nanometres *Nature* **397** 129–34.
79. Hell S. W. 1997 Improving the resolution in far-field fluorescence microscopy by point-spread-function engineering. *Topics in Fluorescence Spectroscopy* (vol. 5) ed. J. Lakowicz (Oxford: Plenum).
80. Bailey B., Farkas D. L., Taylor D. L. and Lanni F. 1993 Enhancement of axial resolution in fluorescence microscopy by standing wave excitation *Nature* **366** 44–8.
81. Hell S. W. and Wichmann J. 1994 Breaking the diffraction resolution limit by stimulated emission *Opt. Lett.* **19** 780–2.
82. Egner A., Jakobs S. and Hell S. W. 2002 Fast 100-nm resolution three-dimensional microscope reveals structural plasticity of mitochondria in live yeast *PNAS* **99** 3370–5.
83. Klar T. A. and Hell S. W. 1999 Subdiffraction resolution in far field fluorescence microscopy *Opt. Lett.* **24** 954–6.
84. Goeppert-Meyer M. 1931 Ueber Elementarakte mit zwei Quantenspruengen, Goettinger Dissertation *Ann. Phys.* **9** 273–94.
85. Kaiser W. and Garret C. 1961 Two-photon excitation in CaF_2:Eu^2+ *Phys. Rev. Lett.* **7** 229–31.
86. Denk W., Strickler J. H. and Webb W. W. 1990 Two-photon laser scanning microscope *Science* **248** 73–6.
87. Denk W., Piston D. W. and Webb W. W. 1995 Two-photon molecular excitation in laser-scanning microscopy. *Handbook of Confocal Microscopy* 2nd edn, ed. J. Pawley (New York: Plenum) pp. 445–58.
88. Summers R. G., Piston D. W., Harris K. M. and Morrill J. B. 1996 The orientation of first cleavage in the sea-urchin embryo, *Lytechinus variegatus*, does not specify the axes of bilateral symmetry *Dev. Biol.* **175** 177–83.
89. Yuste R. and Denk W. 1995 Dendritic spines as basic functional units of neuronal integration *Nature* **375** 682–4.
90. So P. T. C., French T., Yu W. M., Berland K. M., Dong C. Y. and Gratton E. 1995 Time resolved fluorescence microscopy using two-photon excitation *Bioimaging* **3** 49–63.
91. Diaspro A. 2001 *Confocal and Two-Photon Microscopy Foundations, Applications and Advances* (New York: Wiley-Liss).
92. Potter S. M., Fraser S. E. and Pine J. 1996 The greatly reduced photodamage of 2-photon microscopy enables extended 3-dimensional time-lapse imaging of living neurons *Scanning* **18** 147.
93. Gryczynski I., Szmaczinski H. and Lakowicz J. R. 1995 On the possibility of calcium imaging using Indo-1 with three photon excitation *Photochem. Photobiol.* **62** 804–8.
94. Hell S. W., Bahlmann K., Schrader M., Soini A., Malak H., Gryczynski I. and Lakowicz J. R. 1996 Three-photon excitation in fluorescence microscopy *J. Biomed. Opt.* **1** 71–4.
95. Wokosin D. L., Centonze V. E., White J. G., Hird S. N., Sepsenwol S., Malcolm G. P. A., Maker G. T. and Ferguson A. I. 1996 Multiple-photon excitation imaging with an all-solid-state laser *Optical Diagnostics of Living Cells and Biofluids, San Jose: SPIE*, vol. 2678, pp. 38–49.
96. Wokosin D. L., Centonze, V. E., Crittenden A. and White J. G. 1995 Three photon excitation of blue emitting fluorophores by laser scanning microscopy *Mol. Biol. Cell* **6** 113a.
97. Maiti S., Shear J. B., Williams R. M., Zipfel W. R. and Webb W. W. 1997 Measuring serotonin in live cells with three-photon excitation *Science* **275** 530–2.
98. Mohler W., Millard A. C. and Campagnola P. J. 2003 Second harmonic generation imaging of endogenous structural proteins *Methods* **29** 97–109.
99. Zumbusch A., Holtom G. R. and Xie X. S. 1999 Vibrational microscopy using coherent anti-Stokes Raman scattering *Phys. Rev. Lett.* **82** 4142.
100. Cheng J. X., Potma E. O. and Xie X. S. 2002 Coherent anti-Stokes Raman scattering correlation spectroscopy: Probing dynamical processes with chemical selectivity *J. Phys. Chem. A* **106** 8561.
101. Cheng J. X., Volkmer A., Book L. D. and Xie X. S. 2001 An epi-detected coherent anti-Stokes Raman scattering (E-CARS) microscope with high spectral resolution and high sensitivity *J. Phys. Chem. B* **105** 1277.

102. Cheng J. X., Volkmer A., Book L. D. and Xie X. S. 2002 Theoretical and experimental characterization of coherent anti-Stokes Raman scattering microscopy *J. Opt. Soc. Am. B* **19** 1363.
103. Cheng J. X., Volkmer A., Book L. D. and Xie X. S. 2002 Multiplex coherent anti-Stokes Raman scattering microspectroscopy and study of lipid vesicles *J. Phys. Chem. B* **106** 8493.
104. Cheng J. X., Book L. D. and Xie X. S. 2001 Polarization coherent anti-Stokes Raman scattering Microscopy *Opt. Lett.* **26** 1341.
105. Cheng J. X., Jia Y. K., Zheng G. and Xie X. S. 2002 Laser-scanning coherent anti-Stokes Raman scattering microscopy and applications to cell biology *J. Biophys.* **83** 502.
106. Potma E. O., Jones D. J., Cheng, J. X., Xie X. S. and Ye J. 2002 High-sensitivity coherent anti-Stokes Raman scattering microscopy with two tightly synchronized picosecond lasers *Opt. Lett.* **27** 1168–72.
107. Volkmer A., Book L. D. and Xie X. S. 2002 Time-resolved coherent anti-Stokes Raman scattering microscopy: Imaging based on Raman free induction decay *Appl. Phys. Lett.* **80** 1505.
108. Volkmer A., Cheng J. X. and Xie X. S. 2001 Vibrational imaging with high sensitivity via epidetected Coherent Anti-Stokes Raman Scattering microscopy *Phys. Rev. Lett.* **87** 23901.
109. Feofanov A. V., Grichine A. I., Shitova L. A., Karmakova T. A., Yakubovskaya R. A., Egret-Charlier M. and Vigny P. 2000 Confocal Raman microspectroscopy and imaging study of theraphthal in living cancer cells *Biophys. J.* **78** 499–512.
110. Otto C., de Grauw C. J., Duindam J. J., Sijtsema N. M. and Greve J. 1997 Applications of micro-Raman imaging in biomedical research *J. Raman. Spectrosc.* **28** 143.
111. Sijtsema N. M., Otto C., Segers-Nolten G. M. J., Verhoeven A. J. and Greve J. 1998 Resonance Raman microspectroscopy of myeloperoxidase and cytochrome b558 in human neutrophilic granulocytes *Biophys. J.* **74** 3250–5.
112. Sijtsema N. M., Tibbbe A. G. J., Segers-Nolten J., Verhoeven A. J., Weening R. S., Greve J. and Otto C. 2000 *Biophys. J.* **78** 2606.
113. Peticolas W. L., Patapoff T. W., Thomas G. A., Postlewait J. and Powell J. W. 1996 Laser Raman microscopy of chromosomes in living eukaryotic cells: DNA polymorphism in vivo *J. Raman Spectrosc.* **27** 571–8.
114. Otto A. 1984 *Light Scattering in Solids* ed. M. Cardona and G. Guentherodt (Berlin: Springer) p. 289.
115. Kneipp K., Kneipp H., Itzkan I., Dasari R. R. and Feld M. S. 2001 *Proc. SPIE—Int. Soc. Opt. Eng.*, vol. 4258, p. 50.
116. Kneipp K., Wang Y., Kneipp H., Perelman L. T., Itzkan I., Dasari R. and Feld M. S. 1997 Single molecule detection using surface-enhanced Raman scattering (SERS) *Phys. Rev. Lett.* **78** 1667–70.
117. Garcia Vidal F. J. and Pendry J. B. 1996 Collective theory for surface enhanced Raman scattering *Phys. Rev. Lett.* **77** 1163–6.
118. Otto A., Mrozek I., Grabhorn H. and Akemann W. 1992 Surface enhanced Raman scattering *J. Phys. Condens. Matter* **4** 1143–212.
119. Kneipp K., Wang Y., Kneipp H., Itzkan I., Dasari R. R. and Feld M. S. 1996 Population pumping of excited vibrational states by spontaneous surface-enhanced Raman scattering *Phys. Rev. Lett.* **76** 2444–7.
120. Hanlon E. B., Manoharan R., Koo T. W., Shafer K. E., Motz J. T., Fitzmaurice M., Kramer J. R., Itzkan I., Dasari R. R. and Feld M. S. 2000 Prospects for in vivo Raman spectroscopy *Phys. Med. Biol.* **45** R1–59.
121. Manfait M., Morjani H. and Nabiev I. 1992 Molecular events on simple living cancer cells as studied by spectrofluorometry and microSERS Raman spectroscopy *J. Cell. Pharmacol.* **3** 120–5.
122. Efrima S. and Bronk B. V. 1998 Silver colloids impreganting or coating bacteria *J. Phys. Chem.* **102** 5947–50.
123. Cotton R. A. and Fletcher J. H. 1992 Bringing soft x-ray contact microscopy to the biologist *Proc. RMS* **27** 77–80.
124. Cotton R. A., Fletcher J. H., Andrews A. J. and Webb C. E. 1992 Dedicated soft x-ray source for contact microscopy. *X-Ray Microscopy III* (Springer Series in Optical Sciences, vol. 67) ed. A. Michette et al. (Berlin: Springer) pp. 209–11.
125. Fletcher J. H. 1993 *DPhil Thesis* University of Oxford.
126. Fletcher J. H., Cotton R. A. and Webb C. E. 1992 Soft x-ray contact microscopy using laser generated plasma sources. *Proc. SPIE Conf.—Soft X-ray Microscopy*, vol. 1741, pp. 142–53.
127. Cotton R. A. 1990 *DPhil Thesis*. University of Oxford.
128. Cefalas A. C., Sarantopoulou E., Kollia Z., Argitis P., Tegou E., Ford T. W., Stead A. D., Danson C. N., Neely D. and Kobe S. 2003 Nanostructured imaging of biological specimens *in vivo* with laser plasma X-ray contact microscopy *Mater. Sci. Eng. C* **23** 105–8.
129. Dickson R. M., Cubitt A. B., Tsien R. Y. and Moerner W. E. 1997 On/off blinking and switching behaviour of single molecules of green fluorescent protein *Nature* **388** 355–8.
130. Edman L., Mets U. and Rigler R. 1996 Conformational transitions monitored for single molecules in solution *Proc. Natl Acad. Sci. USA* **93** 6710–15.
131. Ishii Y. and Yanagida T. 2000 Single-molecule detection in life sciences *Single Mol.* **1** 5–16.
132. Jia Y. W., Talaga D. S., Lau W. L., Lu H. S., DeGrado W. F. and Hochstrasser R. M. 1999 Folding dynamics of single GCN_4 peptides by fluorescence resonant energy transfer confocal microscopy *Chem. Phys.* **247** 69–83.
133. Lu H. P., Xun L. and Xie X. S. 1998 Single-molecule enzymatic dynamics *Science* **282** 1877–82.
134. Sase I., Miyata H., Ishiwata S. and Kinosita K. 1997 Axial rotation of sliding actin filaments revealed by single-fluorophore imaging *Proc. Natl Acad. Sci. USA* **94** 5646–50.
135. Van Oijen A. M., Ketelaars M., Kohler J., Aartsma T. J. and Schmidt J. 1999 Unraveling the electronic structure of individual photosynthetic pigment-protein complexes *Science* **285** 400–2.
136. Weiss S. 1999 Fluorescence spectroscopy of single biomolecules *Science* **283** 1676–83.
137. Harms G. S., Cognet L., Lommerse P. H. M., Blab G. A., Kahr H., Gamsjaeger R., Spaink H., Soldatov N. M., Romanin C. and Schmidt T. 2001 Single-molecule imaging of L-type Ca^{2+} channels in live cells *Biophys. J.* **81** 2639–46.

138. Kamiyaa H., Fujimuraa Y., Matsuokaa I. and Harashima H. 2002 Visualization of intracellular trafficking of exogenous DNA delivered by cationic liposomes *Biochem. Biophys. Res. Commun.* **298** 591–7.
139. Arndt-Jovin D. J., Robert-Nicoud M., Kaufman S. J. and Jovin T. M. 1985 Fluorescence digital microscopy in cell biology *Science* **230** 247–56.
140. Helmchen F. and Waters J. 2002 Ca^{2+} imaging in the mammalian brain in vivo *Eur. J. Pharmacol.* **447** 119–29.
141. Helmchen F., Fee M. S., Tank D. W. and Denk W. 2001 A miniature head-mounted two-photon microscope: High-resolution brain imaging in freely moving animals *Neuron* **31** 903–12.
142. Damaskinos S., Dixon A. E., Ellis K. A. and Diehl-Jones W. L. 1995 Imaging biological speciments with the confocal scanning laser microscope/macroscope *Micron* **26** 493–502.
143. Monk M. 1987 *Mammalian Development: A Practical Approach* (Washington DC: IRL Press).
144. Ashkin A. and Dziedzic J. M. 1987 Optical trapping and manipulation of viruses and bacteria *Science* **235** 1517–20.
145. Ashkin A., Dziedzic J. M. and Yamane T. 1987 Optical trapping and manipulation of single cells using infrared laser beams *Nature* **330** 769–71.
146. Block S. M. 1990 Optical tweezers: A new tool for biophysics. *Noninvasive Techniques in Cell Biology* ed. B. H. Satir (New York: Wiley-Liss) pp. 375–402.
147. Koenig K., Svaasand L., Liu Y., Sonek G., Patrizio P., Tadir Y., Berns M. W. and Tromberg B. J. 1996 Determination of motility forces of human spermatozoa using an 800 nm optical trap *Cell Mol. Biol.* **42** 501–9.
148. Ishii Y., Ishijima A. and Yanagida T. 2001 Single molecule nanomanipulation of biomolecules *Trends Biotechnol.* **19** 211–16.
149. Smith S. B., Cui Y. J. and Bustamante C. 1996 Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules *Science* **271** 795–9.
150. Greulich K. O. 1999 *Micromanipulation by Light in Biology and Medicine* (Basel: Birkhaeuser).
151. Koenig K. 2000 Laser tweezers and multiphoton microscopes in life sciences *Histochem. Cell Biol.* **114** 79–92.
152. Stout A. L. 2001 Detection and characterization of individual intermolecular bonds using optical tweezers *Biophys. J.* **80** 2976–86.
153. Nishiyama M., Muto E., Inoue Y., Yanagida T. and Higuchi H. 2001 Substeps within the 8-nm step of the ATPase cycle of single kinesin molecules *Nat. Cell Biol.* **3** 425–8.
154. Svoboda K. and Block S. M. 1994 Force and velocity measured for single kinesin molecules *Cell* **77** 773–84.
155. Svoboda K., Schmidt C. F., Schnapp B. J. and Block S. M. 1993 Direct observation of kinesin stepping by optical trapping interferometry *Nature* **365** 721–7.
156. Montag M., van der Ven K., Delacretaz G., Rink K. and van der Ven H. 1998 Laser-assisted microdissection of the zona pellucida facilitates polar body biopsy *Fertil. Steril.* **69** 539–42.
157. Neev J., Schiewe M. C., Sung V. W., Kang D., Hezeleger N., Berns M. W. and Tadir Y. 1995 Animal investigations. Assisted hatching in mouse embryos using a noncontact Ho:YSGG laser system *J. Assist. Reprod. Genet.* **12** 288–93.
158. Tirlapur U. K. and Koenig K. 2002 Targeted transfection by femtosecond laser *Nature* **418** 290.
159. Monajembashi S., Cremer C., Cremer T., Wolfrum J. and Greulich K. O. 1986 Microdissection of human chromosomes by a laser beam *Exp. Cell Res.* **167** 262–5.
160. Greulich K. O. and Weber G. 1992 The light microscope on its way from an analytical to a preparative tool *J. Microsc.* **167** 127–51.
161. Koenig K., Becker T. W., Fischer P., Riemann I. and Halbhuber K. J. 1999 Pulse-length dependence of cellular response to intense NIR laser pulses in multiphoton microscopes *Opt. Lett.* **24** 113–15.
162. Koenig K., Riemann I., Fischer P. and Halbhuber K. J. 1999 Intracellular nanosurgery with NIR femtosecond laser pulses *Cell. Mol. Biol.* **45** 195–201.
163. Bass L. S. and Treat M. R. 1995 Laser tissue welding: A comprehensive review of current and future clinical applications *Lasers Surg. Med.* **17** 315–49.
164. Lauto A. 1998 Repair strength dependence on solder protein concentration: A study in laser tissue welding. *Lasers Surg. Med.* **22** 120–5.
165. Shalem S., German A., Barkay N., Moser F. and Katzir A. 1997 Mechanical and optical properties of silver-halide infrared transmitting fibers *Fibre Integr. Opt.* **16** 27–54.
166. Shenfeld O., Eyal O., Goldwasser B. and Katzir A. 1994 Silver halide fiber optic radiometric temperature measurement and control of CO_2 laser-irradiated tissues and application to tissue welding *Laser Surg. Med.* **14** 323–8.
167. Zhang F., Lee G. M. and Jacobson K. 1993 Protein lateral mobility as a reflection of membrane microstructure *BioEssays* **15** 579–88.
168. Finch E. and Augustine G. 1998 Local calcium signalling by inositol-1,4,5-triphosphate in Purkinje cell dendrites *Nature* **396** 753–6.
169. Katz L. C. and Dalva M. B. 1994 Scanning laser photostimulation: A new approach for analysing brain circuits *J. Neurosci. Methods* **54** 205–18.
170. Neher E. and Zucker R. S. 1993 Multiple calcium-dependent processes related to secretion in bovine chromaffin cells *Neuron* **10** 21–30.
171. Robinson I. M., Yamada M., Carrion-Vazquez M., Lennon V. A. and Fernandez J. M. 1996 Specialized release zones in chromaffin cells examined with pulsed-laser imaging *Cell Calcium* **20** 181–201.
172. Walker J. W., Somlyo A. V., Goldman Y. E., Somlyo A. P. and Trentham D. R. 1987 Kinetics of smooth and skeletal muscle activation by laser pulse photolysis of caged inositol, 4,5-trisphosphate *Nature* **327** 249–52.
173. Kaplan J. H., Forbush B. and Hoffman J. F. 1978 Rapid photolytic release of adenosine 5′-triphosphate from a protected analogue: utilization by the Na:K pump of human red blood cell ghosts *Biochemistry* **17** 1929–35.
174. Otis T., Zhang S. and Trussell L. O. 1996 Direct measurement of AMPA receptor desensitization induced by gluta-matergic synaptic transmission *J. Neurosci.* **16** 7496–504.
175. Somlyo A. P., Walker J. W. and Goldman Y. E. 1988 Inositol trisphosphate, calcium and muscle contraction *Phil. Trans. R. Soc. B* **26** 399–414.

176. Karpen J. W., Zimmerman A. L., Stryer L. and Baylor D. A. 1988 Gating kinetics of the cyclic-GMP-activated channel of retinal rods: flash photolysis and voltage-jump studies *Proc. Natl Acad. Sci. USA* **85** 1287–91.
177. Li W., Llopis J., Whitney M., Zlokarnik G. and Tsien R. Y. 1998 Cell-permeant caged InsP3 ester shows that Ca^{2+} spike frequency can optimize gene expression *Nature* **392** 936–41.
178. Parpura V., Basarsky T. A., Liu F., Jeftinija K., Jeftinija S. and Haydon P. G. 1994 Glutamate-mediated astrocyte-neuron signalling *Nature* **369** 744–7.
179. Naebauer M., Ellis-Davies G. C., Kaplan J. H. and Morad M. 1989 Modulation of Ca^{2+} channel selectivity and cardiac contraction by photorelease of Ca^{2+} *Am. J. Physiol.* **256** 916–20.
180. Morad M., Davies N. W., Kaplan J. H. and Lux H. D. 1988 Inactivation and block of calcium channels by photo-released Ca^{2+} in dorsal root ganglion neurons *Science* **241** 842–4.
181. Kamiya H. and Zucker R. S. 1994 Residual Ca^{2+} and short-term synaptic plasticity *Nature* **371** 603–6.
182. Denk W. 1994 Two-photon scanning photochemical microscopy: Mapping ligand-gated ion channel distributions *PNAS* **91** 6629–33.
183. Pettit D. L., Wang S. S., Gee K. R. and Augustine G. J. 1997 Chemical two-photon uncaging: A novel approach to mapping glutamate receptors *Neuron* **19** 465–71.
184. Williams R. M., Piston D. W. and Webb W. W. 1994 2-photon molecular excitation provides intrinsic 3-dimensional resolution for laser based microscopy and microphotochemistry *J. FASEB* **8** 804–13.
185. Kettling U., Koltermann A., Schwille P. and Eigen M. 1998 Real-time enzyme kinetics monitored by dual-color fluorescence cross-correlation spectroscopy *PNAS* **95** 1416–20.
186. Kinjo M., Nishimura G., Koyama T., Mets U. and Rigler R. 1998 Single-molecule analysis of restriction DNA fragments using fluorescence correlation spectroscopy *Anal. Biochem.* **260** 166–72.
187. Koltermann A., Kettling U., Bieschke J., Winkler T. and Eigen M. 1998 Rapid assay processing by integration of dual-color fluorescence cross-correlation spectroscopy: High throughput screening for enzyme activity *PNAS* **95** 1421–6.
188. Schwille P., Oehlenschläger F. and Walter N. G. 1996 Quantitative hybridization kinetics of DNA probes to RNA in solution followed by diffusional fluorescence correlation analysis *Biochemistry* **35** 10182–93.
189. Korlach J., Schwille P., Webb W. W. and Feigenson G. W. 1999 Characterization of lipid bilayer phases by confocal microscopy and fluorescence correlation spectroscopy *PNAS* **96** 8461–6.
190. Brock R., Hink M. A. and Jovin T. M. 1998 Fluorescence correlation microscopy of cells in the presence of autofluorescence *Biophys. J.* **75** 2547–57.
191. Hasler K., Panke O. and Junge W. 1999 On the stator of rotary ATP synthase: the binding strength of subunit δ to $(\alpha\beta)_3$ as determined by fluorescence correlation spectroscopy *Biochemistry* **38** 13759–65.
192. Pack C. G., Aoki K., Taguchi H., Yoshida M., Kinjo M. and Tamura M. 2000 Effect of electrostatic interactions on the binding of charged substrate to GroEL studied by highly sensitive fluorescence correlation spectroscopy *Biochem. Biophys. Res. Commun.* **267** 300–4.
193. Schuler J., Frank J., Trier U., Schafer-Korting M. and Saenger W. 1999 Interaction kinetics of tetramethylrhodamine transferrin with human transferrin receptor studied by fluorescence correlation spectroscopy *Biochemistry* **38** 8402–8.
194. Tjernberg L. O., Pramanik A., Bjorling S., Thyberg P., Thyberg J., Nordstedt C., Berndt K. D., Terenius L. and Rigler R. 1999 Amyloid β-peptide polymerization studied using fluorescence correlation spectroscopy *Chem. Biol.* **6** 53–62.
195. Huang Z. and Thompson N. L. 1996 Imaging fluorescence correlation spectroscopy: Nonuniform IgE distributions on planar membranes *Biophys. J.* **70** 2001–7.
196. Shapiro H. M. 1995 *Practical Flow Cytometry* 3rd edn (New York: Wiley-Liss).
197. Bigos M., Baumgarth N., Jager G. C., Herman O. C., Nozaki T., Stovel R. T., Parks D. R. and Herzenberg L. A. 1999 Nine color eleven parameter immunophenotyping using three laser flow cytometry *Cytometry* **36** 36–45.
198. Roederer M., DeRosa S., Gerstein R., Anderson M., Bigos M., Stovel R., Nozaki T., Parks D. and Herzenberg L. A. 1997 8-Color, 10-parameter flow cytometry to elucidate complex leukocyte heterogeneity *Cytometry* **29** 1–12.
199. Baumgarth N. and Roederer M. 2000 A practical approach to multicolor flow cytometry for immunophenotyping *J. Immunol. Methods* **243** 77–97.
200. Ekins R. and Chu F. W. 1999 Microarrays: their origins and applications *Trends Biotechnol.* **17** 217–18.
201. Blanchard A. P. and Hood L. 1996 Sequence to array: Probing the genome's secrets *Nature Biotechnol.* **14** 1649.
202. Brett D. M., Bruce P. G., Charles H. P. and David C. T. 1998 Direct protein micro-array fabrication using a hydrogel 'stamper' *Langmuir* **14** 3971–5.
203. Chen J. J., Wu R., Yang P. C., Huang J. Y., Sher Y. P., Han M. H., Kao W. C., Lee P. J., Chiu T. F., Chang F., Chu Y. W., Wu C. W. and Peck K. 1998 Profiling expression patterns and isolating differentially expressed genes by cDNA micro-array system with colorimetry detection *Genomics* **51** 313–24.
204. Lockhart D. J. and Winzelerb E. A. 2000 Genomics, gene expression and DNA arrays *Nature* **405** 827–36.
205. Schena M. and Davis R. W. 1999 Genes genomes and chips. *DNA Microarrays: A Practical Approach* ed. M. Schena (Oxford: Oxford University Press).
206. Ekins R. P. and Chu F. W. 1991 Multianalyte microspot immunoassay–microanalytical 'compact disk' of the future *Clin. Chem.* **37** 1955.
207. Ekins R. P. 1989 Multi-analyte immunoassay *J. Pharm. Biomed. Anal.* **7** 155–68.
208. Edwards A. M., Arrowsmith C. H. and des Pallieres B. 2000 Proteomics: New tools for a new era *Mod. Drug Discov.* **3** 34–4.
209. Emili A. Q. and Cagney G. 2000 Large-scale functional analysis using peptide or protein arrays *Nature Biotechnol.* **18** 393–7.
210. Ge H. 2000 UPA, a universal protein array system for quantitative detection of protein–protein, protein–DNA, protein–RNA and protein–ligand interactions *Nucleic Acids Res.* **28** e3.
211. Irving R. A. and Hudson P. J. 2000 Proteins emerge from disarray *Nat. Biotechnol.* **18** 932–3.

212. MacBeath G. and Schreiber S. L. 2000 Printing proteins as micro-arrays for high-throughput function determination *Science* **289** 1760–3.
213. Gourley M. F. and Gourley P. L. 1997 Integration of electro-optical mechanical systems and medicine: where are we and where can we go? *Proc. SPIE* **2978** 197–204.
214. Gourley P. L. 1998 Resonant-cavity apparatus for cytometry or particle analysis, US Patent No 5,793,485.
215. Gourley P. L. and McDonald A. E. 1997 Semiconductor microlasers with intracavity microfluidics for biomedical applications *Proc. SPIE* **2978** 186–96.
216. Gourley P. L., Cox P. D., Hendricks J. K., McDonald A. E., Copeland G. C., Sasaki D. Y., Curry M. and Skirboll S. L. 2001 Semiconductor microcavity laser spectroscopy of intracellular protein in human cancer cells *Proc. SPIE* **4265** 113–24.

25

Biomedical Laser Safety

Harry Moseley and Bill Davies

CONTENTS

25.1 Introduction 346
25.2 Laser Accidents 346
  25.2.1 Case Studies 346
25.3 Classes of Lasers in Medical Use 348
  25.3.1 Class 1 349
  25.3.2 Class 1C 349
  25.3.3 Class 1M 349
  25.3.4 Class 2 349
  25.3.5 Class 2M 349
  25.3.6 Class 3R 349
  25.3.7 Class 3B 349
  25.3.8 Class 4 350
25.4 Intense Pulsed Light Sources 350
25.5 Principles of Quality Assurance 350
25.6 Laser Standards and Guidelines 353
  25.6.1 Electrical and Mechanical Construction 353
  25.6.2 Optical Radiation Safety Documents 353
  25.6.3 Optical Beam Specification 355
  25.6.4 General Optical Safety Guidance 355
25.7 Laser Safety Management 355
  25.7.1 Optical Radiation Safety Policy 355
  25.7.2 Laser Protection Adviser 355
  25.7.3 Laser Protection Supervisor 355
  25.7.4 Local Rules 356
  25.7.5 Laser Controlled Area 357
  25.7.6 Authorized Users 357
  25.7.7 Training 357
  25.7.8 Medical Examination 357
  25.7.9 Incident Reports 358
  25.7.10 Equipment 358
  25.7.11 Equipment Modification 358
  25.7.12 Legislative Health and Safety Requirements 358
  25.7.13 Conclusion 358
25.8 Precautions 358
  25.8.1 Hazards to the Eye 358
  25.8.2 Hazards to the Skin 358
  25.8.3 Equipment Features 359
  25.8.4 Room Layout 359
  25.8.5 Protective Eyewear 359
  25.8.6 Risk Assessment 359
25.9 Reflections 359
25.10 Fires 360
25.11 Hazards to Patients 361
  25.11.1 Risk to Eyes 361
  25.11.2 Risk to Skin 361
  25.11.3 Misdirected Beam 361

25.11.4 Endotracheal Tube Ignition ... 361
25.11.5 Carbonization ... 362
25.12 Incidental Hazards ... 362
25.12.1 Laser Plume ... 362
25.12.2 Electrical Danger ... 362
25.12.3 Carcinogenic Dyes ... 362
25.12.4 Fire ... 362
25.13 Conclusion ... 362
References ... 363
Further Reading ... 363

25.1 Introduction

The first report of an eye accident came only nine months after the first practical demonstration of a laser. At the same time, the application of lasers to medicine was also developing apace, and only a few months later, a ruby laser was being used in ophthalmology for photocoagulation. The dream of non-contact precise surgery was to some extent fulfilled by subsequent generations of lasers which soon became commonplace in hospital theatres and outpatient clinics. This is one of the rare situations in which someone is exposed to the direct laser beam, and its destructive properties employed for the benefit of the irradiated individual. This brings peculiar requirements with respect to laser safety. In this chapter some of the basic elements of biomedical laser safety are explored, beginning with a discussion of what happens when things go wrong.

25.2 Laser Accidents

A laser accident is essentially a laser safety failure. To be fair, the key question is whether the accident was reasonably foreseeable. The thinking behind a risk assessment approach to safety is to consider what might reasonably be expected to go wrong before it does go wrong. Inevitably, accidents will happen, but a rational approach to laser safety will reduce to a minimum the chances of an accident occurring. It is useful to apply the 'ALARA' principle, according to which the risk should be 'as low as reasonably achievable'. It is worthwhile noting that ALARA is borrowed from ionizing radiation protection where the dose is kept as low as reasonably achievable. This is based on the belief that for ionization radiation there is no threshold. This is not considered to be the case for exposure to a laser beam. Nonetheless, it is useful to apply ALARA to minimizing risk, rather than dose.

25.2.1 Case Studies

There are well-documented cases of incidents involving lasers but it is likely that these are only the tip of the iceberg. There are probably several thousand laser incidents in total allowing for near misses and non-reporting. It should be noted that the accounts given later are based on statements made by the victim and are therefore susceptible to recall error (deliberate or unintentional).

A dramatic account of a laser injury was given by Dr. David Decker who was himself the victim [1]. He was partially blinded by a reflection from what in his opinion was a relatively weak Nd:YAG laser beam. Retinal damage resulted from a 6mJ, 10 ns laser pulse. He was not wearing protective goggles at the time although they were available in the laboratory. It is interesting to note that he did not like to wear goggles because they caused tunnel vision, they became fogged, and they were very uncomfortable after several hours in the laboratory.

This is from Decker's own account 'when the beam struck my eye I heard a distinct popping sound, caused by a laser-induced explosion in the back of my eyeball. My vision was obscured almost immediately by streams of blood floating in the vitreous humour and by what appeared to be particulate matter suspended in the vitreous humour. It was like viewing the world in a round fish bowl full of glycerol into which a quart of blood and a handful of black pepper have been partially mixed. There was local pain within a few minutes of the accident but it did not become excruciating. The most immediate response after such an accident is horror. As a Vietnam war veteran, I have seen several terrible scenes of human carnage, but none affected me more than viewing the world through my blood-filled eyeball. In the aftermath of the accident I went into shock, as is typical in personal injury accidents'. Decker was relatively fortunate since the beam struck the retina between the fovea and the optic nerve. He has been left with a blind spot which he described as being much like having a finger placed over one's field of vision.

Two lessons come out of this incident. One is for the operator to use protective eye wear irrespective of the discomfort involved. The other is for the manufacturer. In the specification of eye wear, user comfort is important. If protective eye wear is comfortable then it is more likely that the user will wear it and this should also be borne in mind when a decision is made to purchase eye wear.

In another reported incident, a laser servicing engineer was working on an Nd:YAG laser making adjustments through an opening in the top of the equipment. In this case goggles were being worn. The engineer was carrying out adjustments on the equipment during beam alignment and this required him to lean over the equipment with his head bent so that he could examine the position of the beam spot on a hand-held target. As he was doing this, his protective glasses slid off his nose. He heard what sounded like a snap in his left eye and saw an immediate bright after-image which lasted for about 20 min. He was left with a blind spot in his left eye which was still evident after 6 weeks.

The safety lesson to draw from this incident is in many ways more difficult since the victim actually did put on laser goggles and, unfortunately, they slipped off. A simple lesson is to ensure that eye wear is fitted securely but that is stating the obvious. Perhaps what is more helpful is to realize that this accident probably took place because the engineer was concentrating on the job at hand, i.e. seeing the spot on the target so that he could align the laser precisely. If any lesson can be drawn from this it is that whatever intricate procedure we are involved in it is essential that we keep in mind the safety implications. Neither our own safety nor anyone else's should be compromised through our concentration on the laser procedure.

A more unusual laser injury occurred in Germany. The source was a laser range finder in an American tank, and the victim was a civilian who was sitting in his car watching the tank, which was on manoeuvres. When the beam from the laser range finder hit his eye, the motorist sustained damage to both eyes. Hopefully, there are not too many similarities between the way lasers are used in hospitals and how they are used by the military. Nonetheless, it is still constructive to consider this case.

It really points to the fact that whoever is firing the laser is responsible for its safe use. This is true in the hospital context as much as elsewhere. The person firing the laser must take responsibility and ensure that all persons present are suitably protected and set the correct atmosphere for safe working practices.

As discussed previously, there are many incidents which are not reported in the open scientific literature. An example from personal experience occurred when a Ho:YAG laser was being used in an ENT procedure. During use, there was a loud crack, described as sounding 'like a whiplash', from the vicinity of the treatment site. The exposure was immediately terminated and it was noted that the aiming beam emerged from the side of the fibre a few millimetres from the distal tip. Treatment continued with a different fibre. Fortunately, there was no harm to the patient as a result of the incident. On examination of the fibre it was seen that the aiming beam emerged from both the distal tip and a hole in the metal sheath some 5 mm from the end. When the Ho:YAG treatment beam was fired down the fibre in the laboratory, there were a few bright flashes from the side of the fibre and within seconds a flame emerged about 5 mm from the tip (Figure 25.1).

The procedure had been carried out with a new fibre. It is worth stressing the value of checking the whole length of fibre prior to use. This should be checked visually and then checked using the aiming beam since any break or discontinuity in the fibre will show up as an area of increased localized brightness. It is also worth emphasizing the need to be alert to events which may be taking place at somewhere other than the target site.

An incident occurred in which a registrar in a hospital suffered retinal damage while using an argon laser for retinal photocoagulation. This procedure was carried out using a slit lamp. In normal operation, when the foot switch was depressed, a protective filter would drop in front of the viewing optics to protect the operator prior to release of the main beam. On this occasion, however, the main beam was fired without the protective filter being in place. It was subsequently found that the fault was mostly due to poor maintenance.

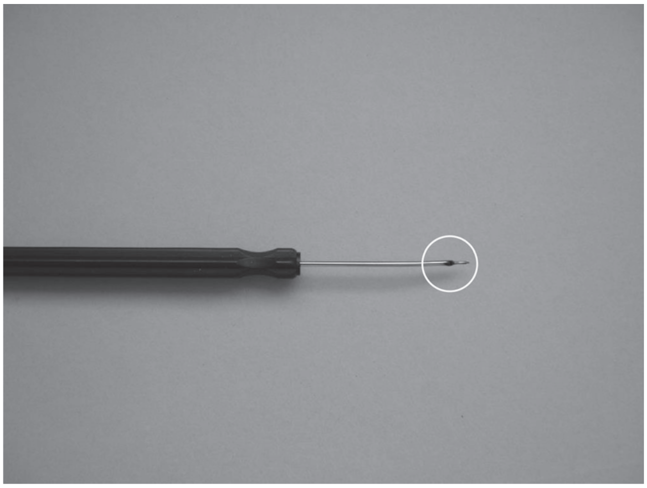


FIGURE 25.1 Fibre used for HoYAG laser ENT application. The flame emerged from the charred area 5 mm from the tip.

The lesson from this is very straightforward which is that, for lasers to operate safely and effectively, there must be regular preventative maintenance carried out.

While considering laser accidents, it should be remembered that there have been cases in which severe damage, sometimes resulting in death, has occurred to patients during laser procedures when the endo-tracheal tube has been ignited by the laser beam. This can occur in ENT applications particularly when the CO_2 laser is used but a case has also been reported of endo-tracheal tube ignition by the Nd:YAG laser. Laser hazards are not simply theoretical possibilities but, if due care is not taken, then they can and do occur.

If the laser is being used anywhere near the endotracheal tube (ETT), then either a laser-safe tube should be used or a non-intubation procedure carried out.

The authors have given evidence in numerous criminal cases involving deliberate exposure of individuals to light from a laser pointer. Cases of actual damage to the eye have now been reported, with current estimates of 150 cases of foveal damage in children in the United Kingdom. In the case of aircraft illumination, the principle hazard is distraction of the pilot, which is potentially serious, if she is involved in a critical operation requiring concentration, e.g. during landing. More than 1400 incidents of lasers being directed at aircraft have been reported in 2015, according to the UK Civil Aviation Authority.

In this chapter, we have chosen to examine closely a few accident situations and analyze these in order to learn how to reduce the risk of causing injury from a laser incident. Another useful resource is the database of laser accidents maintained by Rockwell Laser Industries Inc. This lists 2659 cases but, unfortunately, the most recent incident recorded is in the year 2010. Moreover, incidents involving biomedical laser applications are not identified. The single highest cause of failure is "aircraft illuminated", with 1331 cases (50%) being in this category. The next cause of failure is "negligence", with 463 cases (17%). This is followed by 196 cases of "fibre optic or tip breakage" (7%). These latter two categories are relevant in the biomedical field and are potentially serious.

From examining all the various laser cases which have been reported and also from considering others, from the authors' own experience, some clear safety messages are apparent.

- Accidents are more likely when pulsed lasers are in use since the normal blink-reflex time does not confer any protection.
- Usually no protective eye wear or incorrect eye wear was being worn.
- Many accidents occur during technical adjustment of the laser.
- High-voltage systems are lethal. It is not just the laser beam that can cause damage.
- There is often a lack of awareness of the danger among some operators.
- Lasers in close proximity to anaesthetic gases can be a lethal combination.

25.3 Classes of Lasers in Medical Use

Manufacturers of medical laser equipment are required to assign a hazard classification to each product. This provides guidance on the potential hazards that may be associated with the coherent or quasi-coherent radiation emitted by the system. It does not include consideration of either the hazards associated with any optical pumping process if included, associated electrical power or mechanical hazards. However, users can normally expect that the potential for electrical hazards will rise with classification number.

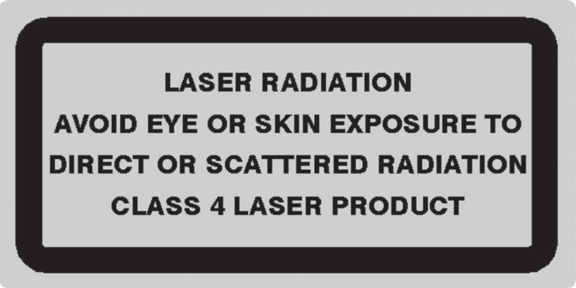
The assigned class is based upon the accessible emission limits (AEL) and should be clearly visible on a label on the exterior of the medical laser (Figure 25.2).

Guidance for manufacturers on the measurement of accessible emission and the subsequent rules for classification are provided in contemporary laser and optical safety standards.

Classes begin with Class 1, the least hazardous, and end with Class 4 the most hazardous.

Solutions to aid the reduction of optical radiation hazard associated with each Class and the engineering requirements to be provided by the manufacturer are set out in Table 25.1.

Unlike industrial lasers the intention of medical laser systems is to deliberately expose tissue either to remove or alter tissue as



LASER RADIATION
AVOID EYE OR SKIN EXPOSURE TO
DIRECT OR SCATTERED RADIATION
CLASS 4 LASER PRODUCT

FIGURE 25.2 IEC label used by manufacturer to indicate classification of laser system.

TABLE 25.1

Engineering Controls Provided by the Manufacturer on a Medical Laser System

	Classification							
Requirements	**Class 1**	**Class 1C**	**Class 1M**	**Class 2**	**Class 2M**	**Class 3R**	**Class 3B**	**Class 4**
Protective Housing	Required for each laser product: medical laser output possible only through the laser aperture. Skin contact device required for Class 1C							
Safety Interlock on Housing	Designed to prevent panel removal until accessible emission values are below Class 3R					Designed to prevent removal of panel until accessible emission values are below that for Class 3B		
Remote control	Not required						Permits easy addition of external interlock although this is not recommended for normal medical use	
Key control	Not required						Laser inoperative when key is removed. Primary engineering means of preventing unauthorized use of laser	
Emission warning device						Gives audible or visible warning when the medical laser-working beam is in a ready or armed state capable of being fired		
Attenuator	Not required						Provides a temporary means of blocking the working laser beam	
Location of controls	Not required					Controls located so that there is no danger of exposure to AEL above classes 1 and 2 except from the laser aperture		
Scanning	Scan failure should not cause product to exceed intended irradiance or radiance to tissue							

Adapted from Table D.2. IEC 60825-1, Edition 1.2 , 2001-08.

in surgery or to stimulate a biological response. The exposure of target tissue and thus the potential exposure of any other accessible normal tissues will be far greater than the maximum permissible exposure (MPE) levels published by ICNIRP.[1] The published MPE levels incorporate generous safety margins of at least one tenth of the experimental ED_{50} levels.[2]

25.3.1 Class 1

Apart from a few small laser diodes sold for acupuncture point stimulation there are very few medical lasers in this class. These lasers are considered safe under all foreseeable circumstances, even intra-beam viewing, Although it should be emphasized that deliberate intra-beam viewing of any laser is not recommended! Typical AEL for continuous wave (CW) lasers are 10mW in the range 1.4–4 µm but much lower for blue light (40 µW). Photochemical effects that are considered to be cumulative give rise to the low AELs for blue light and shorter wavelengths. There are a number of diagnostic medical laboratory systems utilizing Class 3B or Class 4 lasers for cell size analysis that carry a Class 1 designation. This is due to the fact that these more hazardous lasers are totally enclosed within the system during normal operation. Special precautions may be necessary when covers are removed for maintenance or calibration.

25.3.2 Class 1C

Class IC applies to laser products specifically for use in contact with the skin or non-ocular tissue. The irradiance or radiant exposure levels may exceed the skin MPE as necessary for the intended clinical procedure. Engineering means are employed to ensure that accessible emission is stopped or reduced below the AEL of Class 1 when removed from contact with the skin or other non-ocular tissue. These devices may be mechanical or electrical switches that detect skin contact.

25.3.3 Class 1M

M signifies that these are laser beams that are potentially hazardous if magnifying optics are used to collect a larger fraction of the light than that which can be collected by the unaided pupil. This would apply to very divergent or large collimated sources. The latter is not used in medical applications but some laser diode and bare fibre optic sources are very divergent and may fall into this category.

25.3.4 Class 2

This classification only applies to visible radiation between 400 and 700nm. There are many low-power laser alignment systems used in diagnostic imaging to aid proper positioning of the patient. Generally CW, they utilize He-Ne (red and green lines), red laser diodes and frequency doubled Nd:YAGs (green line). Most of the lasers have power outputs in the range 1– 3 mW. The majority of applications use cylindrical lenses to generate a line on the skin further reducing the irradiance at the eye. The CW limit for Class 2 is normally 1mW and many of the green line lasers, which are preferred because of their lower scatter in tissue giving a clearer line on the skin, have output power less than this value. Autonomic responses such as blinking and head aversion response are taken into account in defining the MPE for visible laser beams and care must be exercised if these responses may be absent or compromised in the patient due to medication or absence of a lens (aphakes).

25.3.5 Class 2M

The classification applies to visible radiation as in Class 2 with similar AEL limits but takes account of the two situations that could occur with large collimated beams and very divergent beams where suitable collection optics, i.e. magnification, may increase the retinal irradiance or radiance.

25.3.6 Class 3R

Class 3R applies to visible lasers between 400 and 700 nm with AELs up to five times the Class 1 limit allowing up to 5 mW of continuous power. Direct intra-beam viewing is potentially hazardous but the total exposure takes account of any reductions brought about assumed by aversion responses. Many medical alignment lasers have powers in this range but the irradiance at the eye is reduced by line generation in a cylindrical lens. It is noticeable that patients will not necessarily turn their heads away from bright lights. Class 3R also covers invisible sources in the ranges 180–400 nm and 700 nm–1 mm. There are no current medical lasers laser uses in these wavelength bands producing these irradiance or radiance levels to tissue.

25.3.7 Class 3B

Allowing CW powers up to 500 mW for wavelengths >315 nm means that the majority of lasers used for LLLT (Low light laser therapy) fall in this class. LLLT is used to invoke an immunological response from tissue and typical optical sources are low power laser, laser diode, light emitting diode (LED) or a combination of these sources. The number of sources varies from a single LED or laser diode to clusters of more than 60 sources. Most of these sources have a classification of 3B or lower. Currently LED sources are regarded as not being capable of thermal injury to the eye[3]; however, developments in diode emitter and hybrid device technology such as the vertical cavity surface-emitting laser may produce a radiance of more than 100 $W.cm^{-2}.sr^{-1}$ (100 s average). Direct ocular exposure may be hazardous, even taking into consideration the effect of any aversion response. Diffuse radiation from a Lambertian reflector is not usually a problem with all lasers up to and including Class 3B.

---

[1] ICNIRP [International Committee on non-Ionising Radiation Protection; Revision of guidelines for laser radiation of wavelengths between 400 nm and 1.4 µm: Health Physics, 79(4):431–440:2000.

[2] Laser Safety: Tools and training; 2nd Ed. Ken Barat ed; 2014. ISBN 9781466581371.

[3] ICNIRP statement on light emitting diodes (LEDS) and laser diodes: implication for hazard assessment: Health Physics, Volume 78, No 6, pp. 744–752, 2000.

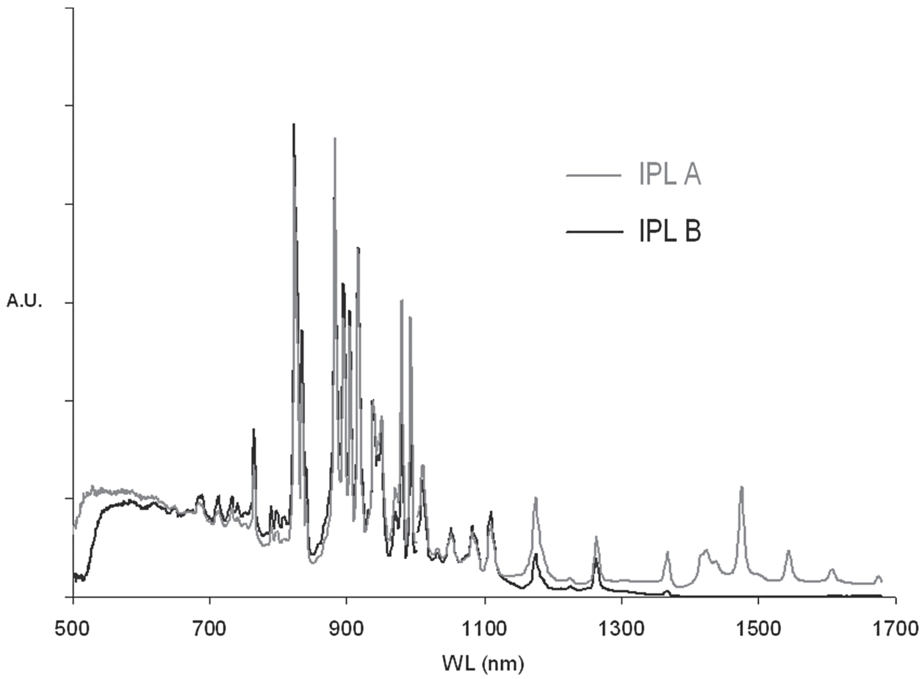


FIGURE 25.3 Calibrated spectral emission from two IPLs. IPL B contained a water filter to block infrared emission. IPL A did not have water filter and so emitted more infrared radiation. (From Ewan Eadie, Studies to Target Optical Radiation Dosimetry Needs in Healthcare, Thesis presented for the degree of Doctor of Philosophy, University of Dundee, 2009.)

25.3.8 Class 4

All lasers with AELs greater than Class 3B fall into Class 4. This includes the majority of surgical and cosmetic lasers. Both direct and diffusely reflected radiation is considered potentially hazardous. There is no upper limit to the irradiance or radiance produced by Class 4 lasers. Users could be dealing with a 1 W diode laser for photodynamic therapy or a 150 W CO_2 laser. The management of these lasers within a healthcare facility requires careful consideration if accidents are to be avoided.

25.4 Intense Pulsed Light Sources

Many cosmetic and aesthetic procedures[4] are now undertaken with polychromatic light sources. These pulsed light sources normally comprise a water- or air-cooled xenon lamp with optical filters used to adapt the bandwidth of the radiation to a specific clinical procedure. Characteristic of these devices are:

1. Polychromatic: range between (475–1100)nm
2. High energy density: up to 60J cm^{-2}
3. Large spot size, usually rectangular: up to 10 cm^2
4. Variable pulse output: µs to ms; this can have complex pulse sequences
5. High-divergence optical output. Up to 50° (0.87 radians)

While the high divergence values reduce the distance over which a hazard exists the bright flash can cause temporary flash blindness over several metres.

Potential hazards exist for both skin and eye exposures close to the emitting surface and a comprehensive rationale for ocular hazard assessment should consider the retinal thermal hazard, blue light hazard (BLH), and the Infrared radiation hazard to the cornea.[5,6] The retinal hazard assessments would be normally undertaken at 200 mm, the shortest distance that the human eye can be expected to focus.

Unlike lasers, IPLs emit light over a wide bandwidth and the shape of the spectrum is not usually available from the manufacturer. However, calibrated data, traceable to national measurement laboratories, have been published and these show, for example, the significant effect of infra-red filtering of the spectrum (Figure 25.3).

25.5 Principles of Quality Assurance

A new medical laser should produce an optical beam output that conforms to specification. Throughout the life and use of the laser changes within the laser system may alter properties of the beam that is delivered to tissue or may increase the risk of associated problems such as electric shock.

---

[4] Goldberg DJ, J. Clinical and Aesthetic Dermatology; June 2012, vol5 (6); Current Trends in Pulsed Light.

[5] Eadie E., Miller P., Goodman T., Moseley H. Assessment of the Optical Radiation Hazard from a Home-Use Intense Pulsed Light (IPL) Source;Lasers Surg Med 2009;41:534–539.

[6] IECTR 60825-9; Safety of Laser products Part 9-Compilation of maximum permissible exposure to incoherent optical radiation.

Quality assurance procedures, based on risk,[7] should be employed to ensure that delivered beam changes at tissue are maintained within an acceptable tolerance and that the mechanics of the whole system remains safe.

Figure 25.4 illustrates a simple process that sets up a procedure called "Quality Assurance" to examine how the medical laser system meets the original specification. The management issues relate to assessing the risks involved from potential changes and deciding when to check critical parameters and who is best placed to carry out these checks.

Parameters that may change over time and can affect the overall safety include:

- treatment beam: beam size and uniformity, power or energy
- beam delivery: articulated arms, fibres, scanners
- electrical safety: cables, foot-switches
- mechanical: enclosure, counterweights, wheels

Important treatment beam parameters are

- beam irradiance, $W{\cdot}m^{-2}$ or energy density, $J{\cdot}m^{-2}$ at tissue target site
- beam shape and uniformity
- beam pulse width
- wavelength

Significant changes in these parameters may affect the outcome for the patient.

What is defined as a significant change will depend upon the treatment protocol. A change of ±20% in output power (tolerance currently set by IEC 60601-2-22) of a CO_2 laser used for polyp ablation may well be tolerable but a 5% change in the output of a laser used in some aesthetic procedures may be considered a significant error.

In many operating circumstances the surgeon will gauge the desired power or energy levels by visual feedback from the ablated tissue.

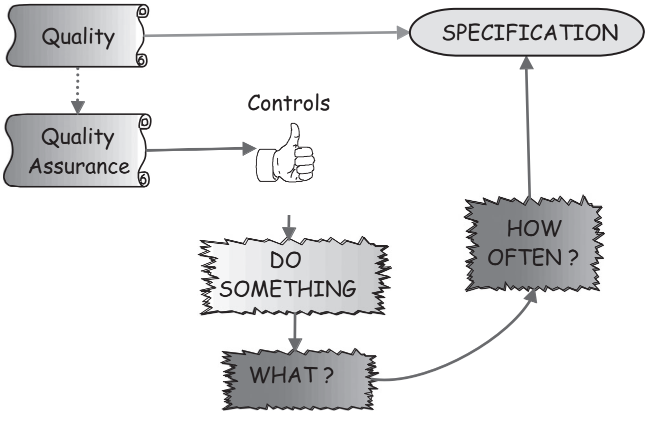


FIGURE 25.4 QA feedback loop.

Measurement of irradiance and radiance in the operating department is unlikely to produce accuracies of better than 10%. Many current medical lasers have a built in independent power or energy calibration that allows the user to measure the output at the distal end of the treatment delivery system. This feature is particularly useful when a long fibre optic is employed for therapy beam delivery. Care must be taken to ensure that the meter pod is clean, as any residue or dirt will lead to an underestimate in the output causing the actual output to be increased.

In surgical dissection using small spot sizes, the power is needed at the middle of the spot with a sharply defined area (Figure 25.5). This image is obtained for a visible beam with a relatively sophisticated camera-based imaging system. A simple guide can be quickly and inexpensively obtained by making a low-power burn on thermal paper or a piece of wood.

The shape and uniformity of the beam and spot mode is important when large spot sizes are used. Larger spot sizes up to 10 mm are employed for such procedures are hair removal or skin resurfacing and it is useful to quickly estimate the approximate flatness of the irradiance or intensity. Thermal paper can give a quick qualitative indicator of laser beam performance as illustrated in the image of a scanned laser beam (Figure 25.6).

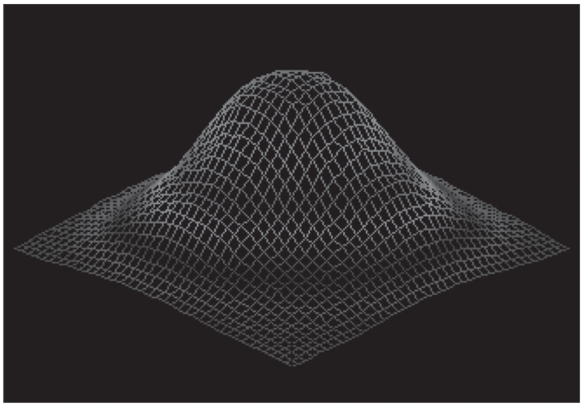


FIGURE 25.5 A smoothly varying 2-D image of a medical laser beam showing a typical Gaussian shape. Approximately 86% of the energy is contained in an envelope of radius $1/e^2$ from the central peak. (Image produced with a Coherent Lasercam II system.)

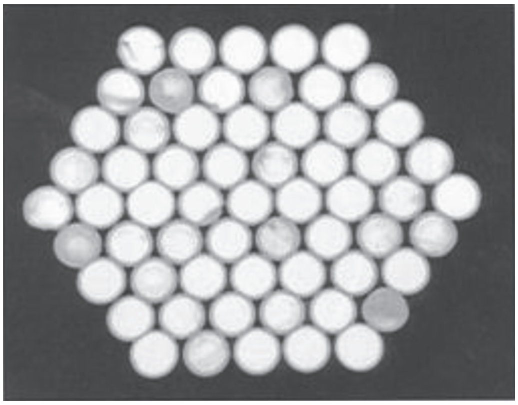


FIGURE 25.6 Burns on thermal paper produced by a scanned ruby laser. (Image courtesy Dr M Kiernan.)

---

[7] BS EN ISO14971:2012, Medical devices. Application of risk management to medical devices.

This is compared with a real-time image from the Lasercam (Figure 25.7) where more details are shown regarding beam edge variations.

With invisible beams such as those produced by a range of infrared lasers, e.g. Er:YAG, CO_2 it is extremely important to know where the beam is going to hit the tissue. Where visible aiming beams are employed in articulated arm delivery systems the coincidence of the aiming beam and the invisible beam should be checked at the beginning of each treatment session. Charring a small piece of wood can readily demonstrate any significant misalignment (Figure 25.8).

By contrast, Figure 25.9 shows a Q-switched NdYAG laser beam which was misaligned as a result of a knock to the articulated arm. This emphasizes the importance of undertaking QA checks.

Laser beams in medical applications can vary from pulses as short as 10^{-15} s (femtosecond) to a CW. Longer pulse lengths can be checked relatively easily using a fast detector and an oscilloscope. Some laser power metres will also display pulse length. Pulse length is defined at the half power points with respect to the peak but is a parameter that does not usually need to be checked on a day-to-day basis. The measurements of pulse widths of ultra-short pulses, ps-fs, are best left to the manufacturer! Occasionally pulse changes in ophthalmology Nd: YAG laser used for posterior capsulotomy procedures are seen (~5 ns) and these will prevent the production of plasma with the consequent loss of cutting ability.

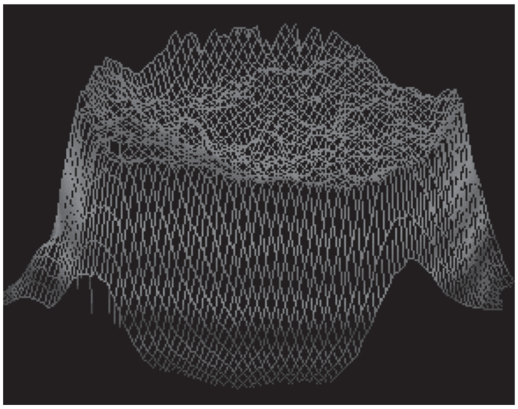


FIGURE 25.7 Image of a nominally flat-topped pulse showing significant variations of energy across the pulse. (Image courtesy of Spiracon©.)

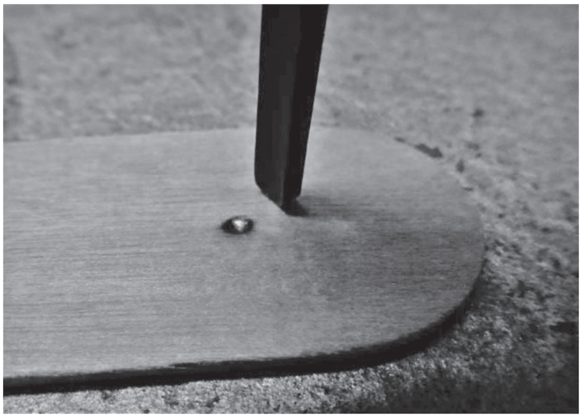


FIGURE 25.8 Image of a CO_2 laser beam burn on a wooden spatula. The red aiming beam is also shown. The black pointer is used to indicate the focal distance of the lens to the tissue target.

The wavelength of the laser operation is unlikely to be a significant variable in most medical applications. Changes of tens of nanometres that may be produced by temperature shift in diode lasers or degradation broadening in the cavity optics are generally tolerable where there are no steeply rising edges in the tissue absorption curve. A spectrophotometer with reasonable resolution is necessary to follow any changes in wavelength and this technique will fall outside normal QA usage except where tuned lasers are employed in medical research applications.

Other quality assurance issues relate to the electrical or mechanical stability of the laser system. Examining the shape, power/energy and position of the treatment beam spot will pick some of the issues relating to the stability of beam delivery systems. Fibre optic systems exhibit fewer mechanical problems than articulated arms. They generally suffer catastrophic and expensive failures such as a complete fibre break. Fibre ends should be carefull examined with the aid of an eye loupe (beam off of course!) to look for damage to the fibre surface. Dirt or scratches can lead to damage the fibre end at therapy powers leading to unwanted patient burn or fires. Many fibres are delivered in a sterile package for single use only.

All footswitches and emergency off switches, power cables, and sockets need regular checking to pick up any issues such as fraying cables, bare wires, etc. The integrity of the mains power cable is particularly important, as it is an integral part of the electrical protection circuit for the patient and operator. Continuity of earthed conductors and panels should be checked

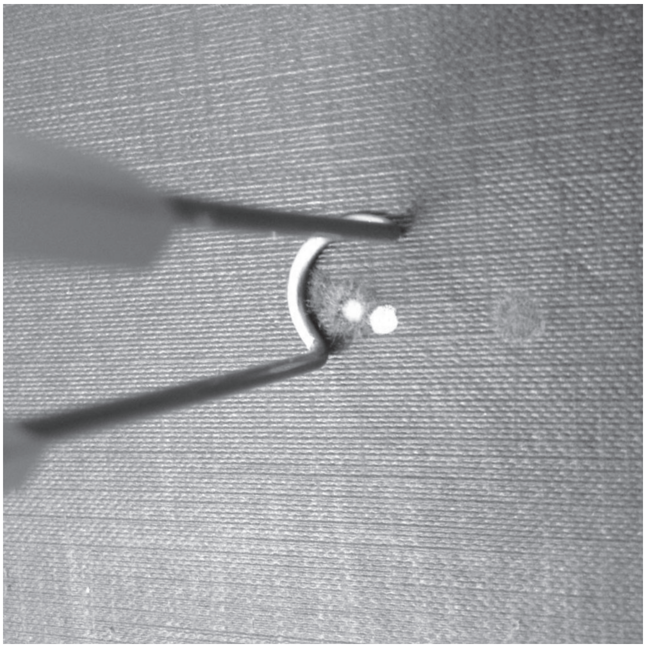


FIGURE 25.9 This shows the red aiming beam from a clinical Q-switched NdYAG laser and the treatment beam clearly separated. In this condition, if the aiming beam were directed on to the target tissue, the treatment beam would be significantly off-target.

TABLE 25.2

International Standards Organizations

Organization	Address	Role
IEC International Electro-Technical Commission	3, Rue De Varembe CH-1211 Geneva 20 Switzerland	Production of international manufacturing standards. Majority of countries subscribe to process.
ISO International Standards Organisation	Casa Postale 56 CH-1211 Geneva 20 Switzerland	Scientific measurement standards for the evaluation of beam optics, component damage, etc.
CEN Comite Europeen de Normalisation	Rue de Stassart, 36 B-1050 Brussels Belgium	European Standards body producing documents linked to European legislation.
ANSI American National Standards Institute	1430Broadway, New York New York 10018	American standards for the US. *Guides are not a legal requirement in all states*

at least annually or following removal of cables or covers for servicing.

An important consideration in the use of most medical laser systems is protective eyewear. This should be checked regularly to ensure that damage has not occurred that could reduce its efficacy.

Like any other piece of medical equipment medical lasers should be part of an appropriate management system that allows for adequate maintenance and calibration at regular intervals.[8] There is a range of simple visual checks that may be undertaken prior to clinical use of the laser and more complex measurements that may be undertaken by suitably qualified personnel at longer intervals. Failure to embark on this process may put both patient and operator at risk of injury.

25.6 Laser Standards and Guidelines

Medical laser standards and guidelines are produced to ensure that consistent information on the safe construction and use of lasers is available to manufacturers, regulatory bodies, and laser users. Some standards also introduce guidance for optical safety practice. The main international organizations producing standards or guidance relevant to medical laser safety are listed in Table 25.2 below:

Inevitably, because of the need to have consensus, international standards may take many years to formulate and publish. From the acceptance of a proposal for new work to publication the International Electrotechnical Commission (IEC) recommends a maximum of 5 years (Figure 25.10).

Medical laser standards and guidance publications address a range of issues

- Electrical and mechanical construction
- Optical radiation safety
- Scientific measurement
- Advice for clinical users

25.6.1 Electrical and Mechanical Construction

Electrical safety for medical equipment is set out generically in a horizontal standard, IEC 60601 Part 1,[9] with particular requirements set out in a vertical standard for the safety of medical laser equipment in Part 2–22.[10] These two publications provide guidance on the essential characteristics of such equipment and specify in some detail the maximum allowable values electrical limits of leakage currents, creepage distances for a range of alternating and direct voltage levels, mechanical stability and acceptable measurement and uncertainty in optical outputs in the case of laser equipment. They constitute a guide for the manufacturer and would normally be transparent to the clinical user. In the USA a similar generic publication[11] exists that is primarily the IEC document although national differences exist that relate to leakage currents, flammability rating and other areas.

25.6.2 Optical Radiation Safety Documents

Optical radiation safety in the wavelength range 180 nm–1 mm is addressed by the IEC through the 60825[12] series of documents. Guidance on medical laser safety is through a combination of part 1 and part 8[13] of this standard.

IEC 60825 Part 1 is drafted by a technical committee of the IEC, namely, TC76.

TC76 addresses a wide area of optical radiation issues through a working group structure. The groups that are of particular interest to medical lasers are

- WG1: fundamental review of biological and physical data to make recommendations/revisions of MPE/AEL and measurement conditions
- WG4: medical laser equipment

---

[8] MHRA2015; Managing Medical Devices Devices.

[9] BS EN 60601-1: 2006+A12:2014 - Medical Electrical Equipment.: General requirements for basic safety and essential performance.

[10] BS EN 60601-2-22:2013: Medical Electrical Equipment Part 2-22: Particular requirements for basic safety and essential performance of surgical, cosmetic, therapeutic and diagnostic laser equipment.

[11] UL 60601-1 2003: Standard for Medical Electrical Equipment, Part 1: General Requirements for Safety:

[12] IEC 60825-1:2014; Part 1: Equipment classification and measurements.

[13] PD IEC TR 60825-8; 2006; Safety of Laser Products: Guidelines for the safe use of laser beams on humans.

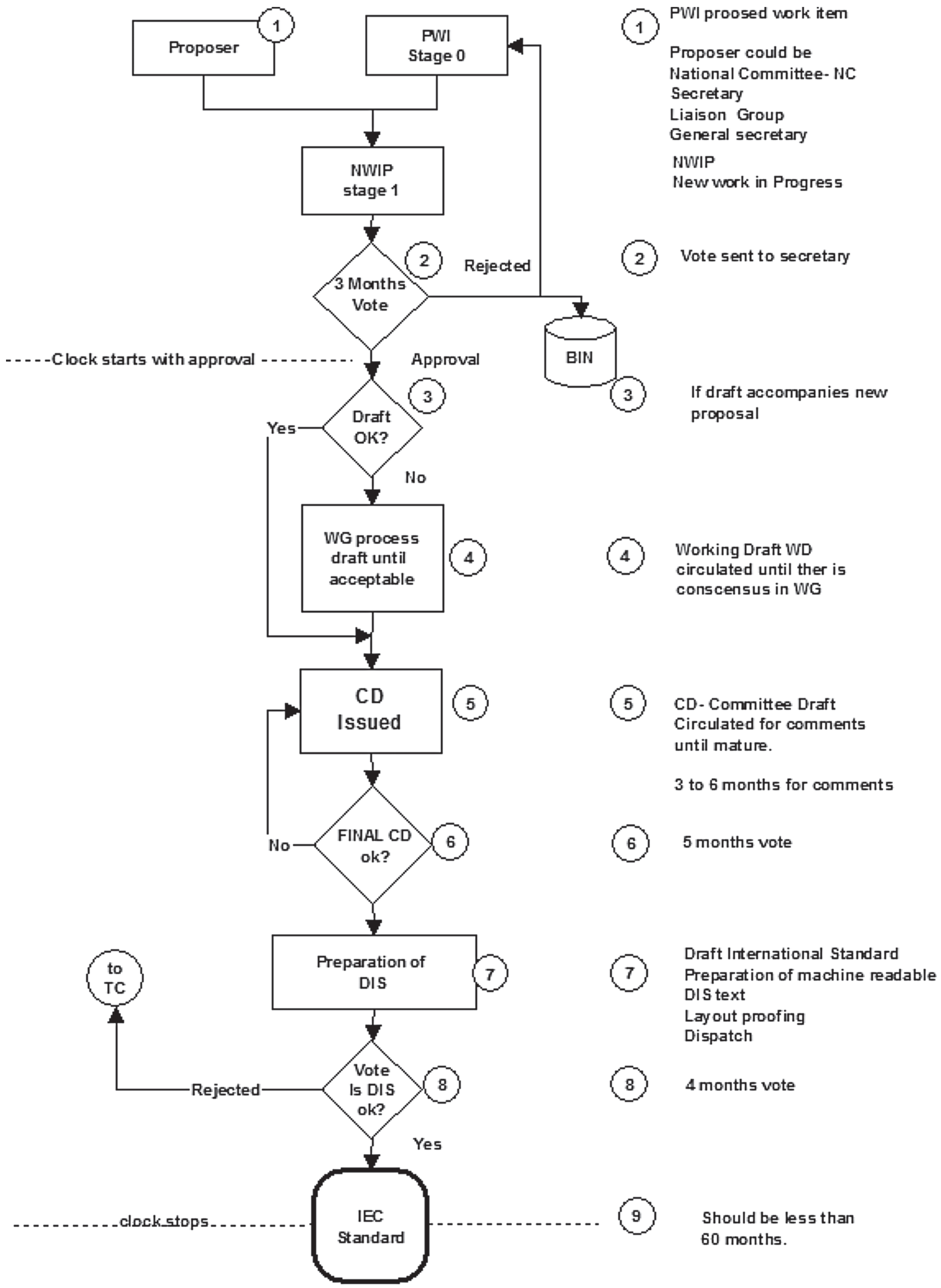

Proposer
1
PWI
Stage 0
NWIP
stage 1
2
3 Months
Vote
Rejected
BIN
-----Clock starts with approval ---------
Approval
3
Draft
OK?
Yes
No
WG process
draft until
acceptable
4
CD
Issued
5
FINAL CD
ok?
6
No
Preparation of
DIS
7
to
TC
Rejected
Vote
Is DIS
ok?
8
Yes
IEC
Standard
-----------clock stops-----------
1
PWI proosed work item
Proposer could be
National Committee- NC
Secretary
Liaison Group
General secretary
NWIP
New work in Progress
2
Vote sent to secretary
3
If draft accompanies new
proposal
4
Working Draft WD
circulated until ther is
conscensus in WG
5
CD- Committee Draft
Circulated for comments
until mature.
3 to 6 months for comments
6
5 months vote
7
Draft International Standard
Preparation of machine readable
DIS text
Layout proofing
Dispatch
8
4 months vote
9
Should be less than
60 months.

FIGURE 25.10 Production of an IEC standard, simplified and abridged from IEC.

Although topics discussed in the other working groups often overlap in subject content WG1 deals with the overarching part 1 standard but in particular looks at new evidence published by ICNIRP[14,15] and that produced from experiments or modelling of laser bioeffects on tissue from existing and new treatment modalities. In response to this evidence it examines the validity of the MPE tables and if necessary recommends changes. These MPE tables provide guidance on the calculation of acceptable levels of occupational exposure to laser radiation for thermal and photobiological effects in the skin and eyes.

WG4 examines areas relating to medical lasers; in particular, it has responsibility for the maintenance of IEC 60601-2-22 and the Technical Report 8[16]: Guidelines for the safe use of

---

[14] ICNIRP Guidelines: on limits of exposure to laser radiation of wavelengths between 180 nm and 1000 μ; Health Physics 105(3): 271–295; 2013.

[15] ICNIRP Guidelines; on the limits of exposure to incoherent visible and infrared radiation; Health Physics 105(1): 74–96; 2013.

[16] IEC TR 60825-8:2006: Guidelines for the safe use of laser beams on humans.

medical laser equipment. The latter report is intended to provide guidance on the practical issues related to clinical use of surgical lasers. Other countries have information of this type embodied into national documents.[17,18]

All the above methods of dealing with laser safety issues are based on deterministic considerations. An alternative technique would be to use probabilistic determination of safety. This approach is well documented in the UK Ministry of Defence document on Military Laser Safety; JSP390.[19]

25.6.3 Optical Beam Specification

Although measurements with respect to the evaluation of classification and MPE currently reside in various parts of the 60825 standards the specification of raw beam parameters such as beam widths, divergence angle, and beam propagation factor are considered in a set of documents produced by ISO.[20,21] These documents are intended to provide manufacturers with an agreed method of specifying beam parameters that could be utilized by clinical users in their safety risk assessment.

25.6.4 General Optical Safety Guidance

Finally, the fourth type of document provides guidance on the practical use of medical lasers and other optical devices. They may give indication of training requirements such as the UK Medical and Healthcare Products Regulatory Agency (MHRA) publication[22] or protocols for clinical treatments as in the ANSI Z136-3 publication.

The implementation of the various medical laser standards varies from useful guidance to mandatory compliance. It is the responsibility of the medical laser user in a particular country[23] to know which regime they are operating under. It is important that the users of medical lasers recognize the need to consult this documentation to help assess the probable risk of injury to patients and staff for each medical procedure.

25.7 Laser Safety Management

Laser safety does not happen by accident. The process has to be managed carefully and effectively. Lasers are used in a variety of circumstances by a wide range of personnel. There has to be a laser safety management structure and a key element of this is the allocation of responsibilities for all staff concerned with the laser. The employer is responsible for health and safety matters and in this respect the employer may be a hospital, university or in the case of a one-man business then the self-employed practitioner is in the eyes of the law the employer. The employer must have a health and safety policy and this is subject to audit by the Health and Safety Inspector.

The Laser Safety Guidance issued by the MHRA in the United Kingdom [5] sets out a structure for laser safety management that has been found to be effective in a wide variety of medical laser applications. It can easily be adapted to other situations and will be used as the model in the present chapter.

The overall and final responsibility for health and safety within an organization rests with the Chief Executive. It is he who is held to be responsible for ensuring that there is adequate resource to produce, implement, and manage effective arrangements for laser safety, but he is not, of course, a laser safety expert.

25.7.1 Optical Radiation Safety Policy

The optical radiation safety policy should define the aims of the senior management and how they implement their commitment to maintaining high standards of safety in the area of optical radiation, i.e. ultraviolet, visible, and infrared radiation, including laser radiation. It should be regularly reviewed and incorporated into the management policies of the organization.

25.7.2 Laser Protection Adviser

It is recommended that the employer should appoint a laser protection advisor (LPA) who will have delegated responsibility for advising the Chief Executive and the board, staff, and the public on laser radiation matters. The LPA should be given adequate information and facilities to perform the work effectively.

The LPA should be knowledgeable in the evaluation of laser hazards and should have responsibility for advising on their control. He/she is expected to carry out a hazard analysis and perform a risk assessment for each laser installation and should ensure that suitable local rules are drawn up and implemented for each Class 3b and Class 4 laser installation. The remit should also extend to intense pulsed lights (IPLs) and LEDs.

The LPA should be consulted on equipment purchase and at the early stages of installation planning. Once the equipment is delivered he or she should ensure that it is ready for use and should set up a scheme of regular safety audits. It is likely that the LPA will be a physicist who has specialized in the area of laser safety. The person appointed as LPA should be able to make valued judgements on the appropriateness of the laser safety arrangements. In the UK, certification schemes are in operation to help ensure the competency of the LPA.

25.7.3 Laser Protection Supervisor

To help with the local supervision within the laser area, a laser protection supervisor (LPS) should be appointed to ensure that the local rules are implemented. The LPS may be a member of staff closely involved with the laser. Each laser installation

---

[17] ANSI: Z136.3-2011, American National Standard for the safe use of lasers in health care facilities.

[18] FDA, Laser Notice No 50, laser Products- conformance with IEC 60825-1 and IEC 60601-2-22; guidance for industry and FDA Staff; June 2007.

[19] Military Laser Safety; UK JSP 390; 2005.

[20] ISO 11146-2:2005. Lasers and laser related equipment—Test methods for laser beam widths, divergence angles and beam propagation ratios—Part 2 General astigmatic beams.

[21] ISO 11146-1:2005; Lasers and laser related equipment—Test methods for laser beam widths, divergence angles and beam propagation ratios—Part 1 Stigmatic and simple astigmatic beams.

[22] Medicines & Healthcare products Regulatory Agency (MHRA); Lasers. Intense light source systems and LEDs- guidance for safe use in medical, surgical, dental and aesthetic practices: September 2015.

[23] ICNIRP Statement on Diagnostic Devices using Non-Ionising radiation: existing regulations and 040.43potential health risks: Health Physics, March 2017, volume 112 number 3, 305–316.

will have a designated LPS whose primary role from the laser safety point of view is to ensure that that particular laser is used safely. The LPS is not expected to be a laser safety expert but if there is any aspect of the use of the laser which is unsatisfactory, then the LPS would be expected to draw this to the attention of the LPA.

The roles of the LPA and LPS are thus clearly differentiated. However, it should be noted that in many organizations there is often one individual who is described as the laser safety officer and this person can be anywhere within the spectrum from laser expert through to record keeper. American and European Standards [4] require that a laser safety officer be appointed to administer a laser safety programme. However, the differentiation of the two roles is now well established in UK hospital safety management and is a well-proven model that could and should be more widely adopted.

25.7.4 Local Rules

The use of lasers should be governed by local rules. These must be applicable to the type of laser which is being used and the particular laser procedures which are being carried out. They must also meet the needs of the local situation. The writing of the local rules requires knowledge of all aspects of laser usage. The rules must be relevant to the particular procedures to be carried out. There is a major difference between a ‘closed application’ in which the beam is enclosed at all times and an ‘open application’ where there is open access to the beam. The latter is clearly a high-risk situation. There are other local issues to be considered such as the ease of access to the laser room by members of the public. All staff involved with the laser have a right to meet and discuss laser safety with the LPA, and accordingly, the local rules should be on display for all to see.

Local rules must also be available for lasers which are introduced temporarily on a trial or demonstration basis. In this case, a more generic approach may be taken to the rules. This inevitably results in protection requirements which are not strictly necessary and would not have a place in local rules for a permanent installation. The reason for this is that they would require more detailed analysis which is not practicable for a short trial period.

There are a number of points which should be addressed within the local rules:

- *Nature of hazards to persons.* It is not appropriate within the local rules to discuss laser hazards in detail. That is, the function of a laser training programme but it is helpful to indicate the sort of hazards which the laser presents. It is useful, for example, to indicate whether both the skin and eyes are at risk if exposed to the laser beam or if the risk is to the eyes only.
- *Controlled and safe access.* The laser should only be used in a designated ‘controlled area’. The controlled area is defined within the local rules. It may simply be the room where the laser is used. The local rules may actually identify more than one area and these are designated controlled areas. Usually there will be a sign on permanent display indicating that the room is a ‘controlled area’. In addition there will be a sign which shows that the laser is actually in use. This may be an illuminated sign.
- *Register of authorized users.* A register should be kept of personnel authorized to operate the laser. Anyone authorized to fire the laser should sign a statement that they have read and understood the local rules. This enforces the principle that the person firing the laser carries the onus of responsibility for its safe use.
- *Nominated key holders.* The laser key should be kept in safe custody and arrangements should be made regarding access to the key.
- *Key holder’s responsibilities.* The key holder should be given explicit written instructions regarding the issuing of the key. Someone who is not an authorized user should be denied access to the laser key.
- *Need for training of all persons.* All staff involved with the laser should receive training at an appropriate level. The LPA should ensure that there is suitable training on laser safety. In addition, anyone carrying out laser treatment should ensure that they are suitably trained in the practical and clinical aspects of the procedure.
- *User’s responsibilities.* It is the person using the laser who is responsible for the safety of all persons present. Although all staff make a contribution to the general level of safety, it is the person firing the laser who is principally responsible.
- *Methods of safe working.* Practical aspects of laser safety should be clearly stated within the local rules. For example, in a room with more than one laser, there should be no confusion as to which laser is on and staff should know which goggles to wear. Another point is to avoid overcrowding. To this end, the number of persons in the room should be kept to a minimum. If the laser beam is directed down a fibre, then the fibre should first be inserted in the patient before the ready switch is enabled. If the laser is being used in an open type of situation, then care should be taken to avoid reflections of the beam from surfaces close to the laser beam. It is very important to ensure that all clamps, lens holders, etc., are secure, and to avoid firing the beam at eye level. Other issues must be considered when the local rules are being written and these depend on the particular procedures being carried out. For example, there is the potentially serious problem of ETT ignition in ENT applications.
- *Pre-use safety checks.* If the laser is to be used to treat a patient, it is good practice to test the laser system before admitting the patient. The laser fibre may be tested visually using the aiming beam. This will show if there is any breakage in the fibre. The therapy beam may be tested simply by firing at a suitable

target and the results may be kept as a record of the equipment performance.

- *Protective eyewear.* The need for personal protective equipment, especially eyewear, should be identified in the local rules and there should also be advice on the type of eyewear required. Generally, everyone in the controlled area would be required to wear protective goggles.
- *Prevention of use by unauthorized persons.* The local rules should clearly state that only authorized operators must use the laser. To avoid misunderstanding, a notice should be placed on the laser stating that it must not be used by an unauthorized person.
- *Normal operating procedures.* These should be written out in a protocol form. This should be based on the manufacturer's own operating instructions.
- *Adverse incident procedures.* The rules should specify the action to be taken in the event of an actual or suspected incident. It is important to have a contingency plan already in place before an accident happens.
- *Contact point for LPA.* A name and telephone number should be given so that the LPA may be contacted at short notice.
- *Limits of the laser controlled area.* The local rules should show clearly the area or areas where the laser may be used. This is usually best done with the aid of a diagram.

25.7.5 Laser Controlled Area

The controlled area is a region within which occupancy and activity are subject to control and supervision. It encompasses an area within which there may be exposure to levels in excess of the MPE levels. The MPE values are defined in the laser standards documents and vary according to the laser wavelength and exposure duration. Access to the area and activity within the area must be controlled as specified in the local rules. In order to designate a controlled area, the LPA should estimate what's called the nominal ocular hazard distance (NOHD). This is the range over which the laser beam represents an ocular hazard. However, it is not advisable to delineate only part of a room as the controlled area. Rather, it is preferable to designate the whole room as a controlled area as this is much easier to implement (Figure 25.11). The advantage of knowing the NOHD is that the LPA can make a judgement based on rational principles as to the level of hazard that exists at the points of entrance to the controlled area. If the NOHD is less than the distance from the laser to the door, then there is no immediate risk to someone who enters the controlled area and there is no need to fit door interlock switches. The door interlock switch is designed to stop the laser if the door is opened and this inevitably interferes with the operation of the laser. In most hospital theatres, fitting a door interlock will introduce greater risk because it interrupts the surgical procedure and so the full implications should be carefully considered when making such a judgement.

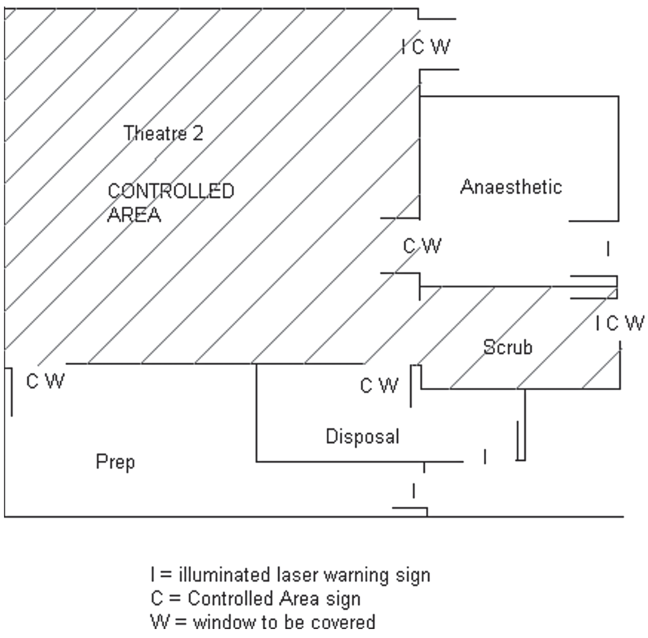


FIGURE 25.11 Designation of laser-controlled area.

It is also helpful to consider the wording to place on signs at the entrance points to the controlled area. Often the words 'caution laser in use' are adequate provided this message is shown only when the laser is powered up. In other situations it may be necessary to instruct staff to wear goggles and this wording should go on the signs. There are occasions when it is necessary to fit a switch to the door, which disables the laser if the door is opened. In this case, the signs should instruct staff not to enter when the laser is on. If the laser is being used in a closed application and the beam is not fired in the air then generally there is no risk on entry to the controlled area.

25.7.6 Authorized Users

All authorized users should sign a statement to indicate that they understand and accept the local rules. This helps to enforce the notion of user responsibility. A register should be kept of staff authorized to fire the laser.

25.7.7 Training

Appropriate training should be given to all who are involved in the use of the laser. It is recommended that everyone who uses a laser on patients should undergo training to acquire the Core of Knowledge which has been jointly agreed by the British Medical Laser Association, the Institute of Physics and Engineering in Medicine, and the Society for Radiological Protection. In addition, practical supervised hands-on training is necessary.

25.7.8 Medical Examination

Regular medical examination is not a necessary part of a laser safety programme. Pre-employment examination, interim, and post-employment examination are of value for medical and legal reasons only. In fact, it has been argued that the light exposure during the slit-lamp examination of the fundus of the eye can in some instances exceed the recommended safe levels and so it is difficult to justify indiscriminate eye examination. On the other hand, if an exposure incident is suspected, then an eye examination should be carried out by an ophthalmologist within 24 h. This will then help to establish whether there has been damage caused to the retina during the laser incident.

25.7.9 Incident Reports

One of the duties of the LPA is to investigate any accidental exposure incident and to produce a report on it. This should clearly show whether the incident was due to equipment failure, user error, or for some other reason. There are always lessons to be learned from a laser accident, and in some cases, this will necessitate changing local rules. As regards equipment faults, within the European community, incidents involving CE marked devices, such as lasers, resulting in serious injury or death must be notified to the appropriate competent authority. In the UK this is the MHRA. Conclusions from such investigations are then disseminated to other EU member states. In the USA, the report is sent to the US Food and Drug Administration.

25.7.10 Equipment

Before purchase, advice should be obtained from the LPA regarding the laser equipment being considered for purchase. The LPA should be able to provide technical knowledge in addition to advice on laser safety matters. It is a requirement within the EU that medical devices be CE marked.

25.7.11 Equipment Modification

Any modifications of laser equipment should be reported to the LPA. It should also be noted that if an organization carries out equipment modification then this may result in the transfer of liability from the manufacturer to the organization.

25.7.12 Legislative Health and Safety Requirements

If the laser safety arrangements are satisfactory there will also be compliance with the appropriate legislation. It is the duty of care which is paramount. Employers are responsible to ensure, as far as it is reasonably practicable, the health and safety of their employees and others who may be affected by their work. Employees themselves also have a duty to take reasonable care of the health and safety of themselves and others and are also expected to comply with the employer's health and safety policy. To this end, it is desirable for the employer to have a laser safety policy. This should specify the responsibilities of the various people involved with the laser, including management. It should clearly state the procedure to be adopted when introducing lasers into the organization.

25.7.13 Conclusion

In conclusion, it is not sufficient to say that laser safety means an absence of accidents. In some cases accidents are simply waiting to happen. Laser safety can be effectively managed.

25.8 Precautions

25.8.1 Hazards to the Eye

The eye is the target organ for laser damage. The optics of the eye cause light and near-infrared radiation to be focused on to the retina to a spot diameter that may be as small as 20 μm. If parallel light is incident on a fully dilated pupil of 7 mm diameter, the increase in flux per unit area from cornea to retina is 10^4. The wavelength band from 400 to 1400 nm is particularly hazardous to the eye because radiation within this range is transmitted to the retina with the potential for causing serious irreparable damage.

Clearly, intra-beam viewing is likely to be more serious than viewing scattered radiation. It is important to ensure that this does not happen. The beam should be enclosed wherever possible and safe systems of work should be in place when the beam is fired in the open. A good safety measure is to avoid firing the beam horizontally at head height. Since specular reflections are equally hazardous, there should be no shiny surface close to the beam. IEC 60825-1 specifies MPE levels at the cornea for ocular exposure to laser radiation [3]. The MPE is defined as that level of laser radiation to which, under normal circumstances, persons may be exposed without suffering adverse effects. There should be no one exposed to a level exceeding the MPE. In the case of an extended source, such as a diffuse reflection or a diode array, the MPE is increased depending on the angular sub-tense at the eye of the observer viewing the source. This allows for the fact that the image of such a source is spread over a larger area of the retina.

Repetitively pulsed or modulated lasers present additional difficulties when assessing the hazard since there is only limited data available and exposure conditions may vary considerably. A structure is presented in IEC 825 [2,3,6]. In this, the most restrictive of three requirements is determined: consider a single pulse within a pulse train; consider the average exposure for a pulse train; finally, consider a single pulse within a pulse train modified by a correction factor that takes account of the number of pulses expected in an exposure.

The distance at which beam irradiance or radiant exposure equals the appropriate corneal MPE is designated the NOHD. This is an extremely useful parameter in deciding the area in which precautions must be taken. NOHD may be calculated if the radiant power, beam divergence, and MPE are known. Beyond the NOHD, no precautions are required. Knowing the NOHD allows an intelligent judgment to be made regarding precautions at doors and windows. It is also helpful when considering room layout. For example, a reflective surface that is positioned beyond the NOHD does not pose a risk provided its position does not change.

25.8.2 Hazards to the Skin

Accidental exposure of the skin may result in a painful laser burn, but this is unlikely to cause serious damage, although a permanent scar may develop. The harm to the skin is much less than to the eye. BSEN 60825 specifies MPE levels for exposure of the skin to laser radiation and these are much higher than for ocular exposure [3]. This is particularly noticeable for visible radiation. For example, between 400 and 700 nm the MPE to the eye for a 1 s exposure is 18 J m^{-2}, compared to 11 000 J m^{-2} for skin. Protection of the skin should not be overlooked but it is secondary to protection of the eyes.

25.8.3 Equipment Features

There are certain safety features that are required by IEC 60825-1 for Class 3B and 4 laser systems [3]. Advice on device features for medical lasers is contained in [5]. Some desirable features are listed here.

Protective covers should be inter-locked to disable the laser if they are removed. Operation of the laser should require the use of a key. There should be a visible indicator that the laser is ready to be fired if the firing mechanism is activated and a warning when laser radiation is being emitted. Beam stops, attenuators, and shutters should be securely fixed and always fail to safety. There should be a remote interlock for connection to a door switch. The laser should also incorporate a target-indicating device, e.g. a non-hazardous aiming beam, unless it is used close to the target. Output power should be measured to an accuracy of 20% for medical systems. There should be a reliable exposure timer and termination system and an emergency laser stop. The equipment should incorporate a 'laser ready' switch and footswitches should be shrouded. Beam transmission systems should be tamper-proof. There should be protective filters and shutters in place if the operator is required to view the irradiation site. Labelling should show the laser classification, maximum output, wavelength, position of output aperture, and suitable warnings. Equipment should also conform to electrical safety requirements and consideration should be given to the BLH.

25.8.4 Room Layout

There are some very simple measures in the room layout that can impact on the safety of the laser. The laser beam should not be directed towards the door, likewise for the windows as appropriate. The beam should be contained as much as possible to reduce the risk of accidental exposure. If the beam is on a bench, all optical components should be securely clamped. Care should be exercised in siting reflective surfaces, as the beam may be misdirected towards them. Avoid firing the beam horizontally at head height because of the risk of accidental eye exposure. Ensure that warning signs are fitted and consider the need for a door interlock taking account of the issues discussed before.

25.8.5 Protective Eyewear

Laser protective eyewear is the last line of defense. Safe systems of work should be in place to ensure that the laser beam does not reach this point. Eyewear is personal protective equipment and there are currently three European standards. EN 207 provides a specification for laser protective eyewear; EN 208 is a specification for eye protectors used for adjustment work on lasers; EN 12254 relates to safety screens used around laser working places [7–9]. Most filters absorb the beam and this causes a temperature rise which leads to thermal damage and burning through the lens. Provided the filter remains intact, the degree of protection is given by the optical density. However, this ceases to be relevant if the damage threshold is exceeded. Eyewear is specified by a scale number that gives the level of protection up to a maximum power or energy density for a particular wavelength. User comfort is also important and this may be influenced by factors such as fit, weight, and ventilation. It is also useful to consider whether wearing the goggles interferes with colour perception and peripheral vision. Eye protection should also be considered when IPL devices are being used and this is covered by a standard issued by British Standards and International Standards Organisation (BS ISO 12609-1 & -2 2013).

25.8.6 Risk Assessment

A hazard is anything that can cause harm; risk is the probability that someone will be harmed by the hazard. There are five steps in risk assessment.

Step 1. *Look for the hazards.* In the current context the laser beam is a hazard. It should be noted what classification it is and its potential for causing harm.

Step 2. *Decide who might be harmed and how.* This may be the person operating the laser, visitors, patient, cleaners, engineers, or members of the public.

Step 3. *Evaluate the risks and decide whether existing precautions are adequate.* With a new installation this will involve appointing a LPA and drawing up local rules.

Step 4. *Record the findings.* It is useful to keep a written record of what was done.

Step 5. *Review the assessment and revise it if necessary.* If there is any significant change in working practice, it is necessary to take account of the new hazard. It is also good practice to review the assessment on a regular basis.

25.9 Reflections

Optical radiation produced by medical lasers covers a wide range of wavelengths from 193 nm (excimer lasers) in the ultraviolet to 10 600 nm (CO_2) in the far-infrared. Reflection is highly dependent on the relationship between the magnitude of the wavelength and the size and optical properties of the materials on which the radiation impinges.

Reflections are either specular, diffuse, or a combination of both (Figure 25.12). Specular reflections obey simple geometric rules and produce reflected radiation paths that are predictable. Specular reflections are utilized in beam delivery systems such as micromanipulators, scanners, and mirror-arm systems to direct the laser beam to the target but back-reflection from the target, off-beam path optical components, operating instruments, and other material in the laser controlled area may pose a safety hazard. Anti-reflection coatings can reduce the first surface reflectivity of around 4% for glass to less than 0.5% for normal incidence. However, it may still be necessary to

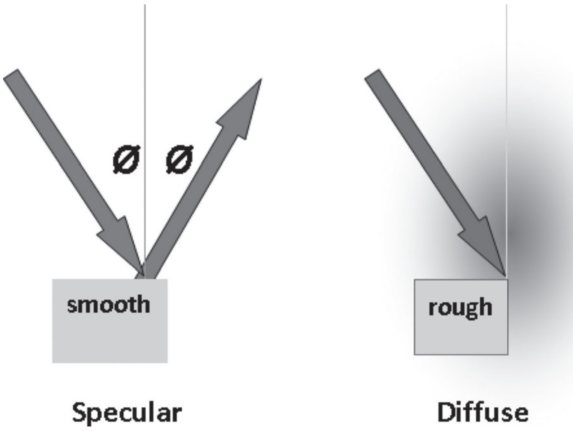


FIGURE 25.12 Specular and diffuse reflections.

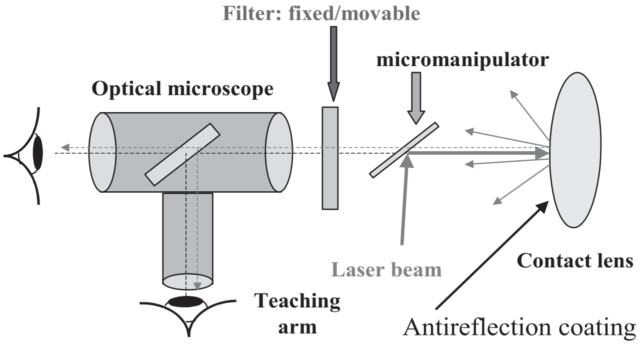


FIGURE 25.13 Anti-reflection coating used in laser eye surgery.

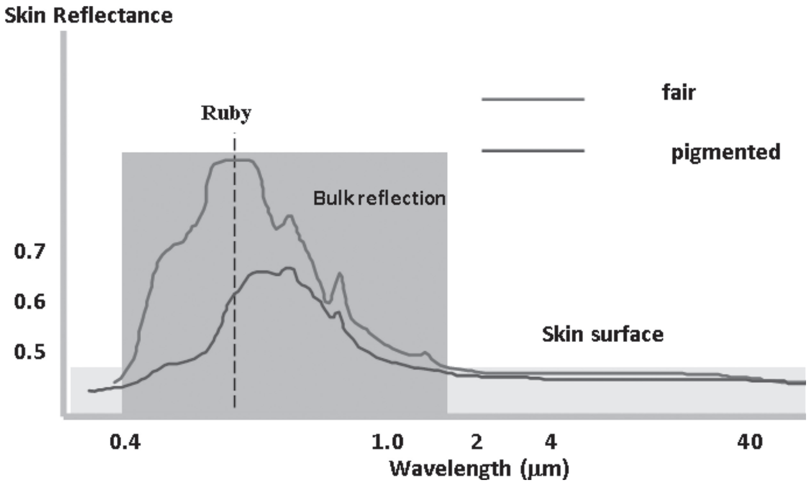


FIGURE 25.14 Reflection of radiation from skin.

incorporate filters into the viewing optical path as shown in Figure 25.13.

In certain situations, with Class 4 lasers, diffuse reflection may pose a hazard but generally with all other classes of lasers diffuse reflection does not present a problem. One application where caution is required is the use of 3B lasers in conjunction with photosensitizing drugs. Diffuse reflected light may produce unwanted reactions in adjacent photosensitized tissue, e.g. treatment of mouth cancers. Reflection from tissues is around a few percent in the ultraviolet and the far infrared where the reflectivity is dominated by refractive index differences between air and the skin. In the visible up to 70% of the light may be reflected. The amount depends on the properties of deeper parts of the skin structures (Figure 25.14).

Within the laser-controlled area the hazards posed by reflected radiation are to the tissue of patients and other persons and unwanted ignition of materials. Reflected laser beam irradiance (W m^{-2}) or radiant energy (J m^{-2}) may be minimized by both controlling the reflectivity of surfaces producing the reflection and considering beam divergence factors in delivery optics. Surface processing of accessory equipment used with CO_2 lasers with sandblasting and fluoropolymer black coatings has been shown to reduce far-infrared reflectivity by up to 1000 times compared to silver finish surfaces (see Figure 25.15) [10].

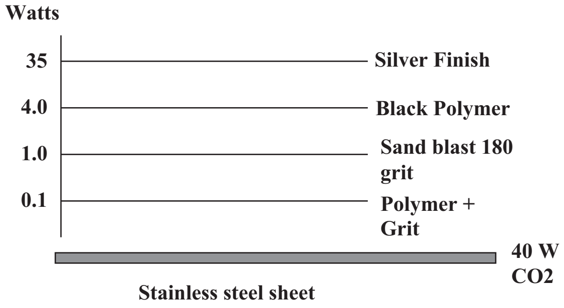


FIGURE 25.15 Reflections from metallic surfaces.

25.10 Fires

Many Class 4 surgical lasers are readily capable of setting fire to materials in the clinical environment. Particular precaution needs to be considered when surgical procedures involve high levels of oxygen, the use of volatile anaesthetics, or inflammable cleaning agents.

Although topical local anaesthesia may leave residues that will flame momentarily the main concerns are related to laser airway surgery and pooling of alcohol-based skin preparations.

The narrow operating space combined with the need to maintain adequate ventilation and gaseous anaesthesia provide the greatest potential for airway fires. High-energy beams can perforate or ignite ETT and anaesthetic gases will support the combustion to produce blowtorch fires in the airway.

Of the current available anaesthetic agents, nitrous oxide, halothane, enflurane, and isoflurane nitrous oxide should be avoided. Older compounds that were problematical such as cyclopropane are rarely used. The higher the thermal conductivity of the anaesthetic agent may be advantageous in preventing fires. Management of the airway gases when anaesthesia is required during laser surgery will reduce but not completely remove the risk of a fire. The incidence of upper airway fires is around a few percent in United States upper airway procedures. Intubation tubes are available in "laser proof" varieties such as Xomed laser shield and Mallinckrott laser flex but none has ideal characteristics without the some potential for igniting. These tubes use stainless steel spirals or metallic powders to reduce the risk of ignition. These tubes are required to meet certain specifications.[24] Conventional tubes such as PVC, red rubber, and silicone may be wrapped metallic tape or protected with metallic backed sponges to shield the flammable material from laser beam contact. Reflection from metallic surfaces may produce unwanted damage in the airway.

Alcohol-based skin preparation have a high flammability and sufficient time to allow evaporation, around 5 min,[25] should be allowed before covering the patient with sterile drapes.

Other sources of fires include hair and GI gases (patients), drapes and dressings, and surgical masks and catheters. It is useful to have a small bowl of sterile saline available to extinguish small fires on the surface of drapes or surgical packing, especially if using a CO_2 laser. Surgical smoke produced from the evaporation of tissue contains tissue debris and organic gases[26] and should be removed as close as possible to the operating target with a filtered smoke evacuator.

25.11 Hazards to Patients

25.11.1 Risk to Eyes

If the eye is the target for treatment, then it is important to ensure that the aiming beam is not hazardous. Also, it is advisable to avoid the use of blue light in the aiming beam because the eye is particularly sensitive to blue light. The characteristics of the laser beam must be precisely controlled, to minimize damage to the healthy eye tissue. There is also a risk of damage to an intra-ocular implant. If the eye is not the target, it should be adequately protected. Wet eye pads or eye shields may be used. If treatment is being given in the vicinity of the eye, tight-fitting eye protectors should be used. For treatments on the eyelids, corneal contact protective shields should be inserted.

25.11.2 Risk to Skin

Serious damage to skin may result in disfiguring scar formation. If this is in a prominent location, it could be very distressing to the patient. Laser treatment to the face, for example removal of port wine stain, resurfacing, or depilation, must be conducted with extreme care. Many laser devices are used in conjunction with some form of skin cooling, which provides protection to the epidermis. This may be done by forced cold air, cryogen spray, cooled window, or a cold ice pack. Although adequate skin cooling is essential, the dangers of over-cooling the skin should also be recognized.

25.11.3 Misdirected Beam

A mis-directed beam will inflict injury on healthy tissue. It may arise from user error, misalignment of the beam, or damage to the delivery fibre. Beam alignment should be checked by firing on a target (e.g. a spatula) prior to use. The delivery fibre should be inspected. This is best done using a low-power aiming beam.

25.11.4 Endotracheal Tube Ignition

Combustion requires a source of ignition, flammable material and an atmosphere that supports combustion. The laser beam is a potent source of ignition. A standard ETT made of red rubber, latex, or PVC may be ignited, and if this occurs in an oxygen-enriched environment supplied by anaesthetic gas mixtures, all the elements are in place for combustion (Figure 25.16). What makes this particularly serious is that the fire burns within the patient's airway. The CO_2 laser in particular and also the Nd:YAG laser have caused such ignition to occur with fatal consequences. Two approaches have been taken to reduce this risk. One is to use a laser-safe ETT. These are either made of metal or silicone rubber with an inner

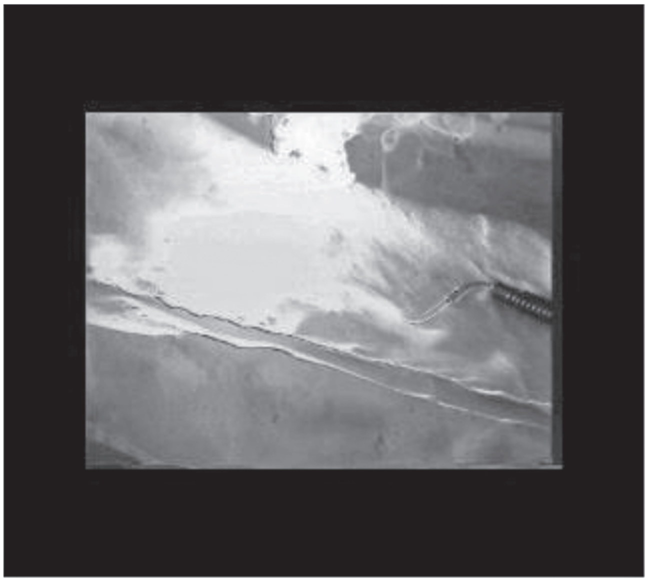


FIGURE 25.16 Dramatic demonstration of a PVC ETT burning fiercely leading to complete destruction under one second. (Reproduced by permission of V Oswal & M Remacle 2002.)

---

[24] BSEN ISO 11990-1;2014: Laser and laser related equipment: Determination of laser resistance of tracheal tubes. Tracheal tube shaft.

[25] New clinical guide to surgical fire prevention (ECRI): Health Devices, 2009; 38(10):330.

[26] Bigony L.; Aorn J., 2007 Dec; 86(6): 1013–20.

aluminium layer. A standard tube may be wrapped in aluminium foil, but this is inferior to the use of a laser-safe tube. An alternative approach is to adopt a tubeless anaesthetic technique. This utilizes jet ventilation and is not only preferable from a laser safety point of view but also provides the surgeon with improved access to the treatment site.

25.11.5 Carbonization

Carbonization takes place during CO_2 or Nd:YAG laser surgery. The charred tissue absorbs the beam producing high localized temperatures that can cause penetrating thermal damage. This may be avoided by the regular removal of charred tissue.

25.12 Incidental Hazards

25.12.1 Laser Plume

During laser surgery airborne contaminants are released into the operating room as a laser plume. The immediate effect of the plume is to obscure the view of the treatment site. It also has a noxious odour which is unpleasant for patient and staff. More serious is the possibility that the laser plume may contain viable particles. Initially it was thought that the high temperature generated during production of the plume was sufficient to destroy cells but it is now thought that this may not be the case at the periphery of the plume. The plume contains particles, gases, carbonized tissue, and biological fragments. Particles of blood, viruses, DNA, and bacteria have been identified in the plume [12,13]. Papillomavirus DNA was demonstrated to be intact. There is the possibility that active viruses may be present in the CO_2 laser plume. Results are somewhat conflicting. Although a laboratory study showed the presence of viable bacteria, there was no active papilloma virus in the plume collected during laryngeal laser treatment [14].

The risk of infection from bacterial- or viral-infected lesions is considered to be low [15]. Likewise transmission risk from malignant tissue is low. Nonetheless, the risk cannot be discounted and steps should be taken to keep this as low as reasonably achievable.

The problem is best dealt with at source. Smoke evacuators with sub-μm filters virtually eliminate the risk to the operator as well as ensuring a good field-of-view. Standard surgical masks do not trap small viral particles, and masks designed to filter viral-sized particles in laser plume are ineffective since the pores become blocked. The smoke evacuator should include a high-efficiency particulate air or ultra-low penetration air (ULPA) viral trap capable of removing particles >0.2 μm with 99.9% efficiency. It should be noted that when using a ULPA filter the suction may be so reduced as to compromise severely the collection efficiency. The best approach is to employ a low-volume high-force pump at the operating site, and a high-volume low-force device to catch the smoke not aspirated by the first suction. If the patient is being treated under local anaesthesia, the laser should be fired during expiration to avoid smoke inhalation by the patient. Since lasing appears to induce localized viral spread, meticulous cleaning of the operating site and surrounding tissue is strongly recommended.

25.12.2 Electrical Danger

Many lasers require three-phase electrical supply and some use potentially lethal capacitor banks. Several deaths have been reported as a result of electrical hazards of laser equipment. The majority are associated with a failure to dump effectively the capacitor banks of high-power lasers. Electrical shock can also arise from careless exposure to the live electrical supply side of modestly powered lasers. So it is important not to overlook high-voltage systems and basic electrical safety.

25.12.3 Carcinogenic Dyes

Some of the dyes used in tunable lasers are surface irritants and some are even carcinogenic. These dyes require periodic changing and must be handled with care. Many of the solvents used in dye lasers facilitate the transfer of the carcinogenic dye through the skin, thereby increasing the risk. Halogen gases in excimer lasers are also highly toxic, even at small concentrations. When these lasers are in use, there must be a rapid evacuation plan in place.

25.12.4 Fire

The laser beam is a concentrated energy source. When absorbed, it can cause a rapid temperature rise. Flammable material should be removed from the vicinity of the laser beam. Gauze placed around the treatment site should be kept moist, and flammable preparations with an alcohol base should be avoided. To prevent fire spreading, there should be easy and quick access to a fire extinguisher (preferably a CO_2 type). When there is a risk of ignition at the treatment site allied with the need for rapid response, e.g. during laryngeal laser surgery, it is a sensible precaution to have a sizeable syringe loaded with sterile saline at hand. It should also be noted that some materials, such as PTFE, may not ignite but produce toxic fumes when heated.

25.13 Conclusion

Lasers have the potential to cause harm and so they must be used carefully. A rational approach to laser safety management is favoured so that the real hazards may be addressed and the risk of accidental exposure reduced as low as reasonably achievable. Such a discussion, however, should not detract from the considerable benefit to be gained from the application of lasers in medicine. In many disease conditions laser therapy is the treatment of choice. Any attempt to use a laser on a patient without regard to safety is counterproductive. The final message is that the safe use of lasers in medicine should be encouraged so that their potential benefits may be fully exploited.

REFERENCES

1. Decker C. D. 1977 An accident victim's view *Laser Focus* **13** 6.
2. IEC 825-5 1998: Safety of Laser Products—Part 5: Manufacturers' check list (BS EN 60825-5).
3. IEC 825-1 2001: Safety of Laser Products—Part 1: Equipment classification, requirements and user's guide (BS EN 60825-1).
4. ANSI: Z136.3-1996, American National Standard for the safe use of lasers in health care facilities.
5. MDA 1995 Guidance on the safe use of lasers in medical and dental practice.
6. IEC 825-8 1999: Safety of Laser Products—Part 8: Guidelines for the safe use of medical laser equipment. IEC, Geneva.
7. EN 207 1994 Specification for filters and equipment used for personal eye protection against laser radiation.
8. EN 208 1994 Specification for personal eye protectors used for adjustment work on lasers and laser systems.
9. EN 12254 1998 Screens for laser working places—safety requirements and testing.
10. Wood R. L., Sliney D. H. and Bayse R. A. 1992 Laser reflections from surgical instruments *Lasers Surg. Med.* **12** 675–8.
11. Oswal V. and Remacle M. (ed.) 2002 *Principles and Practice of Lasers in Otorhinolaryngology and Head and Neck Surgery* (The Hague: Kugler).
12. Byrne P. O., Sisson P. R., Oliver P. D. and Ingham H. R. 1987 Carbon dioxide laser irradiation of bacterial targets *in vitro J. Hosp. Infect.* **9** 265–73.
13. Walker N. P. J., Matthews J. and Newsom S. W. B. 1986 Possible hazards from irradiation with the carbon dioxide laser *Lasers Surg. Med.* **6** 84–6.
14. Abramson A. L., DiLorenzo T. P. and Steinberg B. M. 1990 Is papillomavirus detectable in the plume of laser-treated laryngeal papilloma? *Arch. Otolaryngogol. Head Neck Surg.* **116** 604–7.
15. Ferenczy A., Begeron C. and Richart R. M. 1990 Human papillomavirus DNA in CO_2 laser-generated plume of smoke and its consequences to the surgeon *Obstet. Gynaecol.* **75** 114–18.

FURTHER READING

Apfelberg D. B. (ed.) *Evaluation and Installation of Surgical Laser Systems* (New York: Springer).

Henderson A. R. 1997 *A Guide to Laser Safety* (London: Chapman and Hall).

Hetzel M. R. (ed.) 1995 *Minimally Invasive Techniques in Thoracic Medicine and Surgery* (London: Chapman and Hall).

Moseley H. and Haywood J. K. (ed.) 1987 *Medical Laser Safety* (London: The Institute of Physical Sciences in Medicine).

Moseley H. 1994 Ultraviolet and laser radiation safety *Phys. Med. Biol.* **39** 1765–99.

Moseley H. 1988 *Non-ionizing Radiation* (Bristol: Adam Hilger).

Oswal V. and Remacle M. (ed.) 2002 *Principles and Practice of Lasers in Otorhinolaryngology and Head and Neck Surgery* (The Hague: Kugler).

Puliafito C. A. (ed.) 1996 *Laser Surgery and Medicine: Principles and Practice* (New York: Wiley).

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

26

Laser in Communications: Section Introduction

John Marsh, Subhash C. Singh, and Chunlei Guo

Optical fibre systems have revolutionized telecommunications. First proposed by Kao and Hockham [1] in 1966, the development of optical communication systems presented many challenges. First, low-loss fibres based on glass had to be produced. Second, optical communication systems differ from radio in that the photon energy is much larger and so receivers operate close to the quantum noise limit. The design of the total receiver (detector plus amplifier) is, therefore, critical. Finally, the development of the optical source has presented some of the greatest challenges. The original trial systems used light-emitting diodes, as these could be modulated at MHz frequencies. However, it was the development of room temperature semiconductors lasers that enabled the high-data rate, long-haul systems of today.

The semiconductor laser is ideally suited to optical communication systems because

- it is compact;
- its mode size is well matched to that of the core of single-mode optical fibre;
- it operates at low voltage and current; and
- it can be directly modulated to several tens of GHz

As well as being used as the transmission source, semiconductor lasers are also used to pump optical amplifiers.

Optical devices are becoming integrated with electronics, and this combination has the potential of overcoming some of the electronic bottlenecks associated with moving data on and off VLSI chips. Chapter 26.1 is, therefore, devoted to the development of vertical-cavity surface emitting laser arrays for use with VLSI. Satellite systems are considered in their own Chapter 26.2. A review of modern silicon photonics systems is given in Chapter 26.3. The current technologies of fibre-optic transmission systems are described in Chapter 26.4. Chapter 26.5 describes a recently emerged field of visible light communications and LiFi. In the final Chapter 26.6, different types of optical data storage are reviewed with a focus on optical memory for non von-Neumann architecture.

REFERENCE

1. Kao K C and Hockham G A 1966 Dielectric-fibre surface waveguides for optical frequencies *Proc. IEEE* **113** 1151–8.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

27

Fibre-to-the-Chip: Development of Vertical Cavity Surface-Emitting Laser Arrays Designed for Integration with VLSI Circuits

Ashok Krishnamoorthy, L. M. F. Chirovsky, K. W. Goosen, J. Lopata, and W. S. Hobson

CONTENTS

27.1 Introduction and Background 367
27.2 VCSEL: The P^2I^2 VCSEL Design 368
27.3 First Implementation: The I^2-VCSEL 370
27.4 Design Modifications for High-Speed P^2I^2-VCSELs 376
27.5 Results from the First High-Speed P^2I^2-VCSELs 377
27.6 Conclusions 379
Acknowledgments 380
References 380

27.1 Introduction and Background

It is becoming evident to the builders of large digital routing and cross-connect systems that optical links provide the most efficient and scalable intra-system interconnect to the boards, shelves and bays of their equipment. In such systems, the interconnect technology typically represents a performance bottleneck, as electrical interconnect technologies struggle to keep pace with increasing demand for more connectivity and bandwidth. To address this, a number of parallel optical fibre interconnect modules have been developed and several have appeared in the marketplace (Figure 27.1). Several system vendors of routers and cross-connects have adopted parallel optical links due to their performance (size, weight, power), their span (up to 600 m with graded-index multi-mode fibre), scalability (up to 120 Gbit s^{-1} in a single module) and relatively low cost when compared to alternative telecommunications optical components. In addition to the transport issue, equipment makers have transitioned from bus-based architectures to switched interconnection networks, wherein line-cards residing in a particular shelf or rack of equipment communicate with other line-cards on the same or separate shelf *via* a non-blocking switch. For systems that favour optical transparency at very high line rates and need relatively infrequent reconfiguration of the switch (i.e. certain optical cross-connects), all-optical switches provide a scalable alternative to electrical cross-bar switches. However, for applications that require fast reconfiguration (e.g. packet switching and routing) or aggregation of multiple lower-speed lines, a scalable alternative to electrical switch technology has not yet emerged. Instead, a hybrid approach with optical links and electrical switching appears to be best suited for this application. In this paradigm, it is important to reduce the power and cost of the optical-to-electrical interface, which inevitably calls for the integration of microlasers and detectors to the electrical interface and switching circuits (Figure 27.2). Such technology will ultimately allow low-power, high-bandwidth systems that are based on fibre and surface-normal optical inter-connects for data-transport and silicon VLSI for switching (Figure 27.3). This chapter will focus on the key integration technologies, with an emphasis on the design and fabrication of vertical cavity surface-emitting

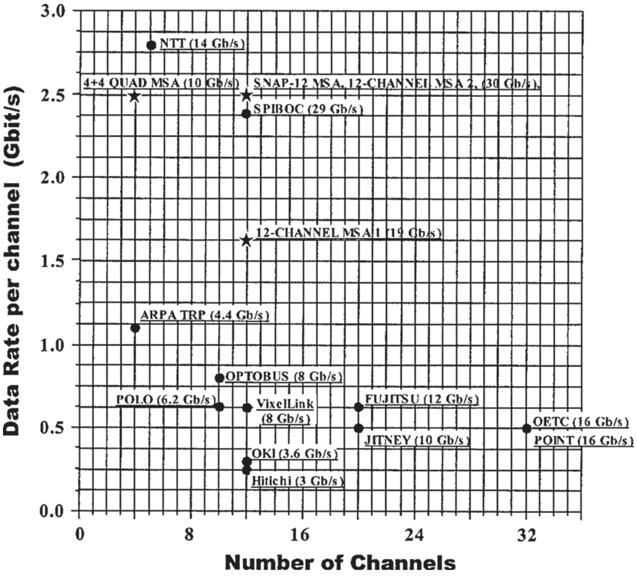


FIGURE 27.1 Progress in 1D parallel optical inter-connect research and development. Projects denoted by a star are now commercial products. New products using 2D array technologies entered the market in 2002 with 36 or more channels operating at 2.5 Gbit s^{-1} per channel and higher.

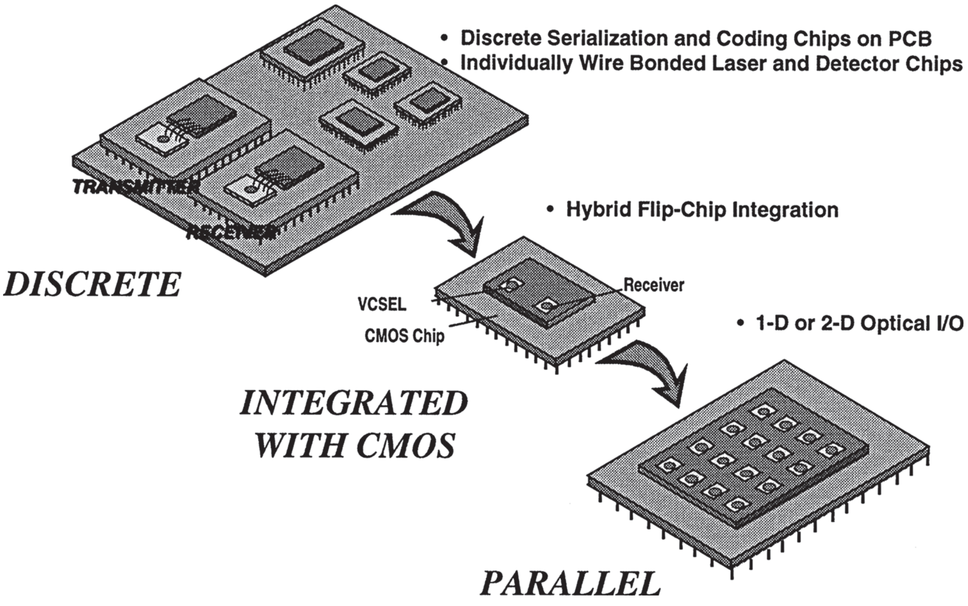

• Discrete Serialization and Coding Chips on PCB
• Individually Wire Bonded Laser and Detector Chips
TRANSMITTER
RECEIVER
DISCRETE
• Hybrid Flip-Chip Integration
VCSEL
CMOS Chip
Receiver
INTEGRATED WITH CMOS
• 1-D or 2-D Optical I/O
PARALLEL

FIGURE 27.2 Integration via flip-chip bonding of VCSELs and detectors to CMOS can provide parallel optical input directly to complex CMOS OE-VLSI switching and processing circuits.

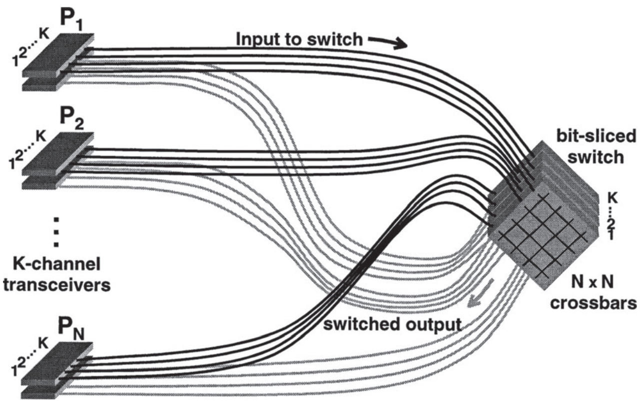


FIGURE 27.3 An optically inter-connected system architecture based on an OE-VLSI switch.

lasers (VCSELs) that are suitable for intimate integration with VLSI electronics.

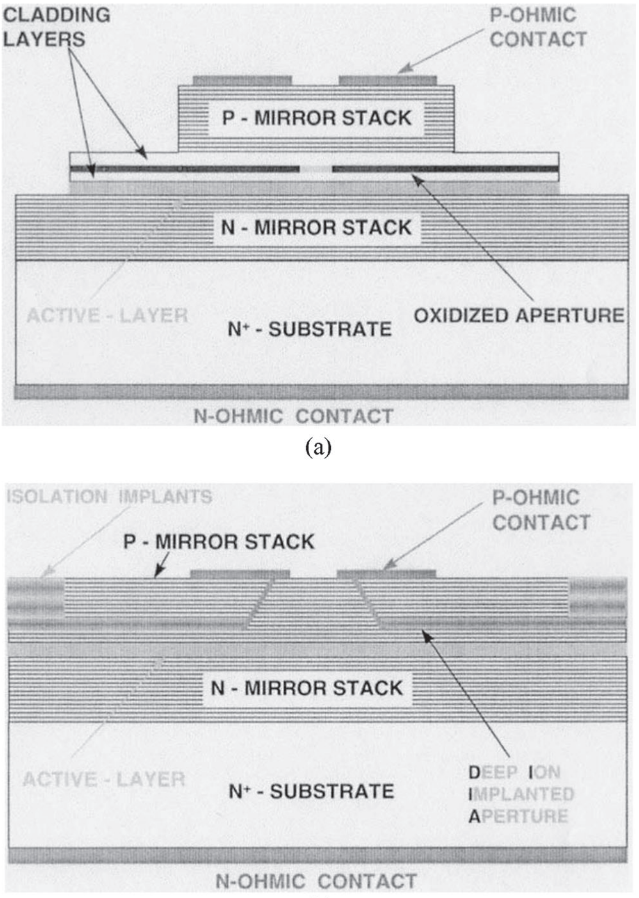
There have been several research efforts around the world whose aim has been to develop what is called 'Optoelectronic VLSI' (OE-VLSI) technology, the intimate integration of photonic devices with VLSI electronics. The general goal is to give the electronics the option of communicating through optical interconnects [1–8]. Most of the early work has focused on using the multiple quantum well (MQW) electro-absorption modulator diode [9,10] as the photonic device, primarily because the MQW diode could serve both as a fast detector and as a modulator, for data reception and for data transmission, respectively. In a monolithic integration approach [11], the MQW diodes were co-fabricated with GaAs field-effect transistors to produce optoelectronic circuits. This technology produced several system demonstrators and was the basis of a multi-project foundry shuttle of monolithic optoelectronic chips. It became obvious that this monolithic approach would have needed a huge development effort to become a VLSI technology, on the scale of the one that produced the silicon CMOS VLSI technology. Such a large effort could not be justified. Instead, attention turned to a hybrid approach, whereby arrays of MQW diodes were flip-chip solder-bump bonded to CMOS chips [12,13], arguably the most advanced electronic VLSI technology. With such chips, encouraging system experiments were performed. Soon thereafter, this hybrid technology produced many more system demonstrators and was the basis of several multi-project foundry shuttles of hybrid optoelectronic chips to research groups around the world [14]. A true OE-VLSI technology was emerging, just as strong arguments developed for the advantages of OE-VLSI technology in general.

In the meantime, many other groups around the world were developing alternative approaches to OE-VLSI. Some of these are described in other chapters of this Handbook, and most of the others are summarized in [15,16]. The challenging problem was to design efficient light sources that could be fabricated and integrated in arrays with the electronics. Because of their power efficiency, large bandwidth and suitability for arrays, VCSELs were the leading candidates for this integration. Because the market for VCSEL-based optical links was also rapidly maturing, VCSELs quickly became the photonic devices of choice for data output, with certain specific design modifications that promoted their ability to be attached to silicon VLSI circuits using an industry-standard flip-chip bonding technique.

27.2 VCSEL: The P^2I^2 VCSEL Design

One of the principal advantages of VCSELs is their high-power conversion efficiency at low power levels, in the 1 mW regime. Previously, only oxide-confined current aperture VCSELs [17,18], using the oxidation of high Al concentration AlGaAs, have achieved threshold currents well below 1 mA, while still maintaining high quantum efficiency. This is because oxide apertures have precise boundaries near the active region and perform beam guiding through an index-guide effect that confines optical beam waists to sizes smaller than all other device features, which minimizes optical loss. The cross section of a generic oxide-confined current aperture VCSEL is illustrated in Figure 27.4a. An issue with such apertures, however, is that it is difficult to control their sizes precisely. The best control of sizes is obtained with photolithographically defined processing steps. Oxide aperture fabrication typically begins with a photolithographic step but the subsequent deep etch (through a thick epitaxial top mirror) and the lateral oxidation procedure can easily introduce significant variations in the final lateral dimensions and, hence, the final electrical and optical characteristics of the device.

There are many possible sources of variation [19]. The errors in the diameter of the aperture prior to oxidation are carried unchanged to the significantly smaller oxide aperture, where they constitute a much larger percentage change in dimensions. The oxidation rate depends heavily on the Al concentration in the AlGaAs and on the thickness of the AlGaAs layer, which are likely to vary across the wafer and from wafer to wafer. The oxidation rate depends on the composition of the surrounding layers, which may also vary. The rate depends on furnace temperature, the water vapour content of the ambient gas and the gas flow dynamics in the oxidation furnace. Lateral oxidation typically proceeds from the boundary of a mesa, which is 10–15 μm away from the aperture centre. It must be controlled to within a few percent, in order to tailor the areas of small apertures (3–6 μm diameters) to within ±10%, which can result in a threshold uniformity of ±10%. Because this control can be difficult to maintain, the variations in the lateral dimensions of an oxide aperture will typically be larger than for a strictly photolithographically defined feature, until an 'oxidation stop' is forthcoming.



(a)

(b)

FIGURE 27.4 Schematic cross section of (a) generic oxide-confined VCSEL; (b) a typical implant-apertured VCSEL.

In principle, these problems can be avoided with photolithographically defined ion-implanted apertures [20]. However, most implanted aperture VCSELs have had even greater troubles. A schematic cross section of one is shown in Figure 27.4b. The difficulties of previous implantation schemes stem principally from the fact that the implantation must penetrate deeply through the top epitaxial mirror, losing their precision and ability to scale to ever-smaller dimensions. That is because the resulting boundaries of the implanted isolation layer become fuzzy due to straggle and rise to the surface of the wafer in an inwardly sloping funnel shape. As a result, the apertures become narrower near the surface than near the active region far below the surface. In order to overcome high series resistance, the opening in the top ohmic metal contact must be several micrometres smaller than the current aperture near the active region. This necessity and the lack of beam confinement by any index-guiding (other than gain guiding) cause more carrier and optical losses in such implant-apertured VCSELs than in oxide-apertured devices. The losses become proportionately greater as apertures are made smaller.

The design approach described here provided an alternate way to obtain the advantages exhibited by oxide apertures, without sacrificing the dimensional accuracy obtainable with photolithography. Note that the VCSELs had to be designed to be amenable to integration with electronic circuits, specifically to the flip-chip solder-bump bonding technique. This meant bottom-emitting VCSELs with the additional design goal of having both p and n contacts as nearly co-planar as possible. In conventional designs, current was injected into a VCSEL cavity through its mirrors (Figure 27.4). This would typically entail one contact to be placed above the VCSEL mesa, and the other at the base, resulting in a large difference in height between contacts (several micrometres). The intra-cavity contact, wherein the active region was directly contacted, was still not conventional but its feasibility had already been proven [21]. With this contacting scheme, the current could be injected directly into the cavity, bypassing the mirror layers on either the p-side or n-side of the device. Dual intra-cavity contacts for both n *and* p contacts would fulfil the design goal best (Figure 27.5), since the vertical separation between the two contacts to each VCSEL could then be less than half a micrometre. With this contacting method came the option of optically optimized mirrors with no doping or grading.

This option was chosen, and was taken one step further, where the top mirror was a deposited dielectric distributed Bragg deflector (dd-DBR). The risk in using a dd-DBR was balanced by several advantages and at least one research group had been using a variation successfully for several years [22].

The alternating layers of a dd-DBR could have a much larger ratio of indices of refraction than any epitaxial DBR. As a result, the spectral bandwidth of a dd-DBR could be much larger than that of epitaxial mirrors (e.g. 300 nm for a dd-DBR *versus* 100 nm for an epitaxial DBR), and the optical penetration depth into a dd-DBR was much shorter. In addition, a desired reflectivity could then be achieved by a dd-DBR with far fewer layer pairs. These three facts translated into the following benefits—much greater tolerance to variations in layer thicknesses, a shorter and so more efficient resonant cavity and a lower profile top mirror that would interfere less with flip-chip bonding.

A somewhat subtler and perhaps even more important advantage of using a dd-DBR as the top mirror of a VCSEL became apparent. This mirror could be deposited and patterned as the last step of the VCSEL fabrication sequence. As a result, processing options for the other device features became available, which would have been impractical with a thick epitaxial mirror in the way. Another important option was that a separate feature on the wafer surface could be formed just under the mirror to confine the diameter of the beam waist throughout the VCSEL structure. This feature could be used to guide the beam as efficiently as possible in the active region, in the current aperture, in the contact openings and in both mirrors. All of these processing steps produce photolithographically defined device features (including the current aperture), thus enhancing the ability to create reproducible yield and performance.

The result was the P^2I^2-VCSEL design strategy, yielding VCSELs processed with photolithographic precision, containing an intra-cavity contacted current guide and an independent optical beam guide [23]. The intra-cavity contacted current guide made the device amenable to flip-chip bonding for intimate integration with electronic circuitry. Processing with photolithographic precision offered the promise of the best reproducibility in fabrication. The independent optical beam guide in combination with the well-defined current guide provided high efficiency. The first implementations of the P^2I^2-VCSEL design strategy were called I^2-VCSELs [23].

27.3 First Implementation: The I^2-VCSEL

In the two older approaches discussed in the previous section, difficulties arose for the same reason. In both cases, critical device features (the guides) were formed deep below the surface, and as a result, these features suffered a loss of precision and/or accuracy in their lateral dimensions. Since the intended lateral dimensions were of the order of 10 μm or less, even a depth of several micrometres could cause the problems discussed previously. In the P^2I^2-VCSEL design approach, the critical features were formed at or near the surface so that photolithographic precision could be accurately transferred to them before being 'buried' under a deposited top mirror. Similar mirrors had been effectively used before [17,22] but were not patterned. As is shown in Figure 27.5a, the generic P^2I^2-VCSEL comprised an undoped substrate, an undoped (optically optimized) epitaxial bottom mirror, the laser cavity, which contained the entire intra-cavity contacted pin structure and a deposited and patterned dielectric top mirror. The cavity region was a few half-wavelengths thick, i.e. no more than a micrometre thick. Five major layers, arranged from the bottom mirror to the surface, constituted the cavity region: a layer of one conductivity type (p or n); an intrinsic active layer; a low doped layer of the opposite conductivity type; a highly doped layer of the opposite conductivity type and a cover layer (undoped). The beam guide was formed in the cover layer, right at the surface. The current guide was formed just below the surface, by contacting the highly doped layers and by creating an apertured isolation region in the low doped layer just above the active layer. The highly doped layers served as dual lateral current injectors sending their respective charge carriers towards the aperture, which then forced the carriers to flow vertically and to recombine in an area of the active layer, approximately equal to the area of the aperture, which was photolithographically defined. Only after all these parts were made was a top dielectric mirror deposited on the surface and patterned to expose the contacts. Care was taken to make sure that the lateral dimensions of the top mirror were large enough to avoid diffractive losses.

For the first implementation, the I^2-VCSEL, the choices were the simplest ones available that satisfied the criteria of the design strategy. The beam guide was a shallow mesa, a small fraction of a wavelength thick and concentric with the current aperture. This mesa caused an abrupt change in the thickness of the cavity at the mesa edge. Light can resonate in such a cavity only if the beam waist becomes small enough for the tails of the Gaussian beam to satisfy the radial boundary conditions for resonance. The situation is analogous to having a small index change, at the same radial position as the edge of the mesa, along the entire length of the cavity, with no mesa. Therefore, this mesa can be called an index guide. Although its effects are not entirely equivalent to those of a typical waveguide, they can dominate over the gain-guiding effects of the current aperture, similar to an oxidized aperture. The apertured isolation layer of the I^2-VCSEL was made with a shallow ion-implantation of a heavy ion. This implant formed precise current apertures in the low p-doped layer under the thin highly conducting lateral injection layer and avoided the problems encountered with deep implants. The two features of index guide and implanted aperture duplicated the effects of a laterally oxidized aperture, with the additional advantage that their effects could be independently optimized. A schematic diagram of the cross section of an I^2-VCSEL is shown in Figure 27.5b. For the purpose of flip-chip solder-bump bonding of I^2-VCSEL arrays to CMOS circuits, gold posts (just slightly taller than the top mirror) were formed on the ohmic contacts. Figure 27.6 is a micrograph of an I^2-VCSEL ready for flip-chip bonding [23].

The very first I^2-VCSELs were top-emitters and so were tested only at wafer level for basic performance properties. They were used to show that (i) the current guide could provide a low-enough series resistance; (ii) the devices would actually lase with patterned dd-DBRs; (iii) the performance of I^2-VCSELs would be comparable to that of oxide-confined VCSELs with similarly sized apertures, as predicted and (iv) the beam guide would work as expected. Figure 27.7 presents typical voltage and differential series resistance *versus*

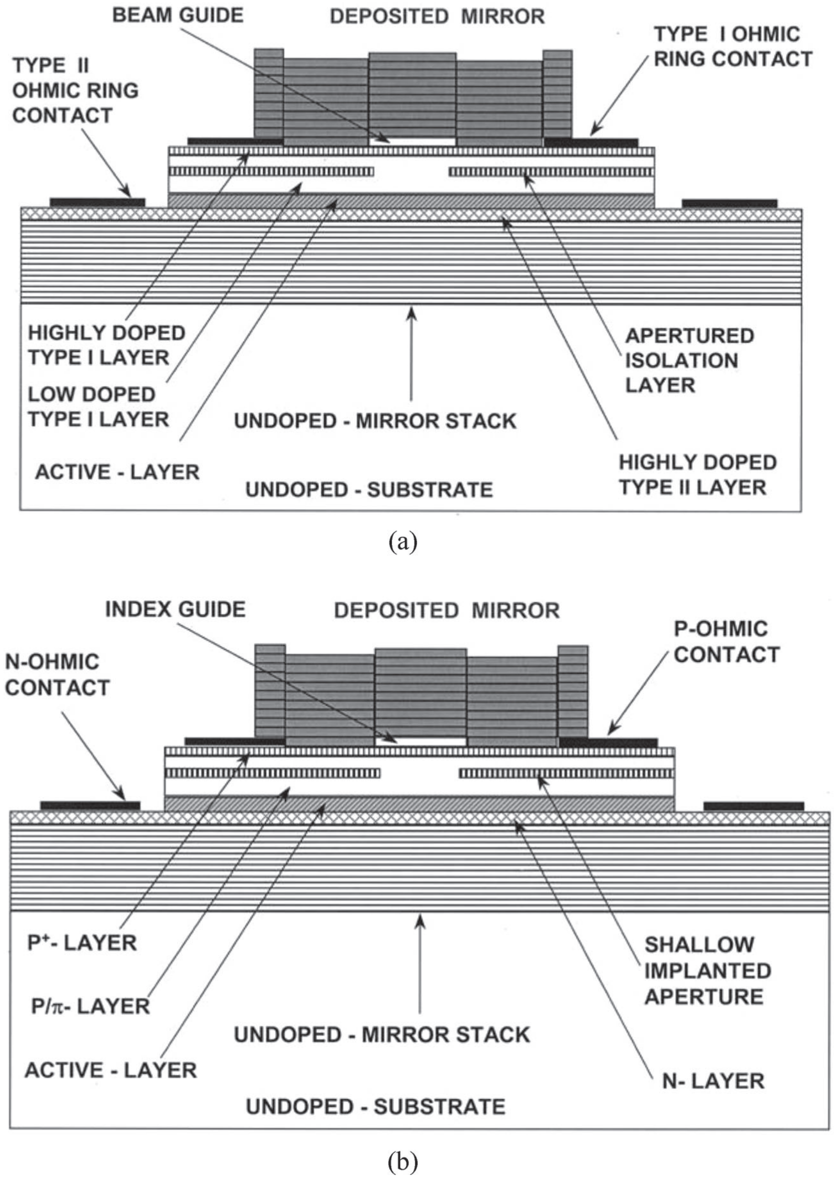

BEAM GUIDE
DEPOSITED MIRROR
TYPE II OHMIC RING CONTACT
TYPE I OHMIC RING CONTACT
HIGHLY DOPED TYPE I LAYER
LOW DOPED TYPE I LAYER
ACTIVE - LAYER
UNDOPED - MIRROR STACK
UNDOPED - SUBSTRATE
APERTURED ISOLATION LAYER
HIGHLY DOPED TYPE II LAYER
(a)
INDEX GUIDE
DEPOSITED MIRROR
N-OHMIC CONTACT
P-OHMIC CONTACT
P+- LAYER
P/π- LAYER
ACTIVE - LAYER
UNDOPED - MIRROR STACK
UNDOPED - SUBSTRATE
SHALLOW IMPLANTED APERTURE
N- LAYER
(b)

FIGURE 27.5 Schematic cross section of (a) a generic P^2I^2-VCSEL; (b) an I^2-VCSEL amenable to flip-chip bonding. The bottom mirror is an AlAs/GaAs quarter-wave stack, the n-layer is ~1000 Å thick and Si-doped to 1E18 cm^{-3}, the active layer comprises three quantum wells and suitable barriers for 980 nm, the p-layer is ~3000 Å thick and Be-doped from 2E17 to 1E18 cm^{-3}, the p$^+$-layer is ~300 Å thick and C-doped to 1E20 cm^{-3}, the index-guiding mesa is a GaAs layer (a fraction of a wavelength thick) atop a thin undoped InGaP stop-etch layer, the contacts are metals commonly used for alloyed contacts, and the deposited mirror is a dielectric quarter-wave stack.

current plots of devices with a 10 μm diameter aperture, proving the excellent properties of the lateral current injection layer. Figure 27.8 shows that the *L–I* curve of a 6 μm aperture I^2-VCSEL (full) was vastly superior to the *L–I* curve for the previously best similarly sized implanted VCSEL (dotted line) and rivalled the *L–I* characteristics of a typical similarly-sized oxide-confined VCSEL [24]. Figure 27.9 shows the richly multiple transverse mode output of a 10 μm device, indicative of a strong index-guided beam confinement that helped improve its *L–I* characteristics. Note that devices with 6 μm diameter apertures already had a single-mode output, with a side-mode suppression ratio of over 30 dB, even when operated at currents ten times the threshold current, far beyond the expected operating current level. This is illustrated in Figure 27.10.

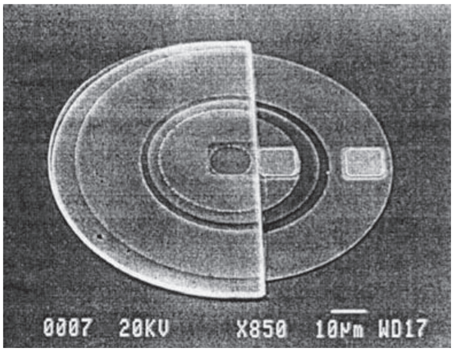

0007 20KV X850 10μm WD17

FIGURE 27.6 A scanning electron micrograph of a completed I^2-VCSEL, ready for flip-chip bonding.

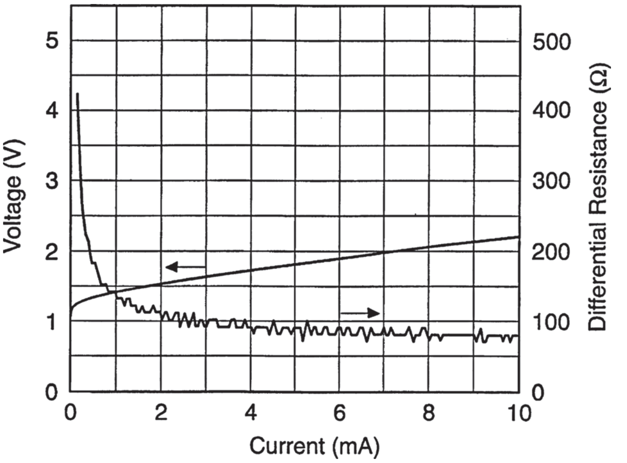


FIGURE 27.7 Electrical characteristics of a typical 10 μm aperture I²-VCSEL, consisting of plots of the voltage *versus* the current and the differential series resistance *versus* the current.

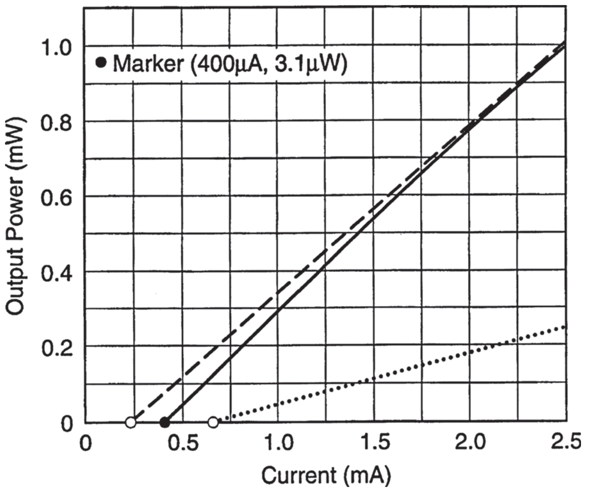


FIGURE 27.8 Plots of the *L–I* curves for a 6 μm aperture I²-VCSEL (full line) the best reported implanted 6 μm aperture VCSEL prior to this work (dotted line) and a typical oxide-confined 6 μm aperture VCSEL made by a potential manufacturer (dashed line).

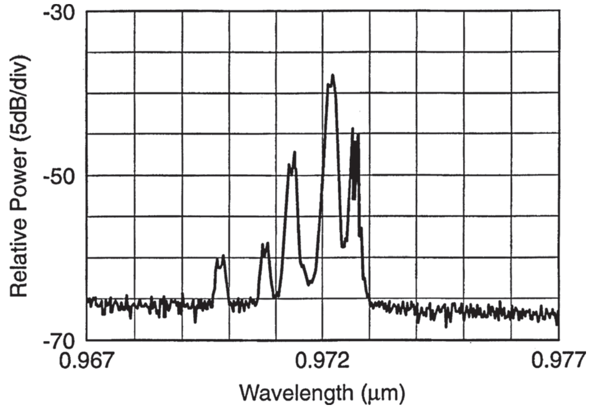


FIGURE 27.9 The multiple-transverse-mode optical spectrum, at 0.1 nm resolution, of the output of a 10 μm aperture I²-VCSEL at 5 mA input, indicating strong index-guided beam confinement.

Since the top-emitting devices had been fabricated as if they were to be flip-chip bonded, the fabrication sequence for the bottom-emitters was similar. Several samples of bottom-emitting I²-VCSELs were made in order to compare their performance to that of the top-emitters. Figure 27.11 shows results from a sample, with many arrays of bottom-emitting devices, where good uniformity over its ~3 cm² area was obtained with characteristics similar to those of the top-emitters. On another sample, the metallization, but not the phase-matching layer, was removed from the top mirrors of some of the VCSELs. For those devices, both mirrors had the same output-level reflectivities and the devices lased bi-directionally. Figure 27.12 shows the *L–I* and the *I–V* curves of a bi-directionally emitting I²-VCSEL and an adjacent bottom-only emitting one (top mirror still metallized). Despite an ~1% output (i.e. loss) per round-trip pass by the light, the bi-directional device had a relatively modest increase in threshold and maintains high slope efficiency, thereby testifying to the high quality of the device material in general. These results were sufficient to attempt the integration of arrays of I²-VCSELs to silicon chips.

The first flip-chip bonding experiments were with 2 × 10 I²-VCSEL arrays and bare-die silicon tester chips. In the beginning, a few VCSELs per array were probed and characterized just prior to the flip-chip bonding and then re-tested just afterwards, through the appropriate contact pads on the silicon tester chips. Figure 27.13 contains the characteristics, taken before and after flip-chip bonding, of one such typical I²-VCSEL. Note that both the *L–I* and the *I–V* responses improved as a result of the bonding. These improvements could be attributed to better electrical and thermal conductivity through the solder-bump contacts. Figure 27.14 shows that the flip-chip-bonded VCSEL, once AR-coated, could have a smooth *L–I* curve, free of kinks.

These 2 × 10 I²-VCSEL arrays were then used to determine whether such devices could be easily integrated and operated with electronic circuitry [25]. Because the lasers operated at a wavelength of ~970 nm, their light was transmitted through

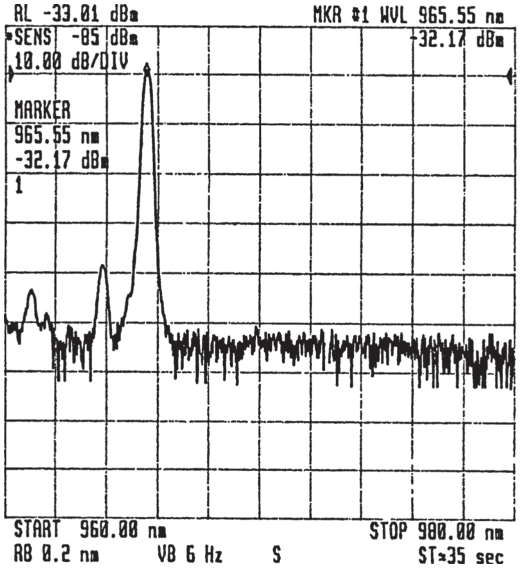


FIGURE 27.10 Single-mode spectrum of a 6 μm aperture I²-VCSEL.

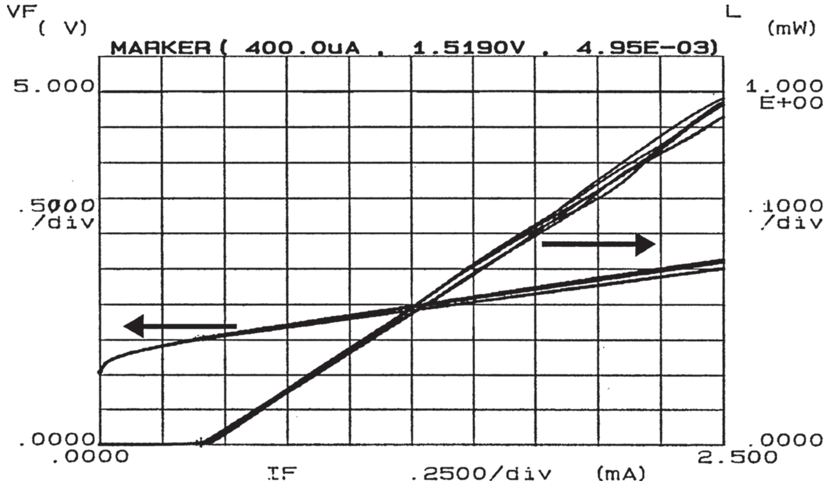


FIGURE 27.11 The *L–I* and *V–I* curves of six randomly chosen (6 µm diameter aperture) bottom-emitting I^2-VCSELs on a sample ~3 cm^2 in area. They show that good uniformity was obtained over the entire sample with a performance comparable to that of top-emitting I^2-VCSELs.

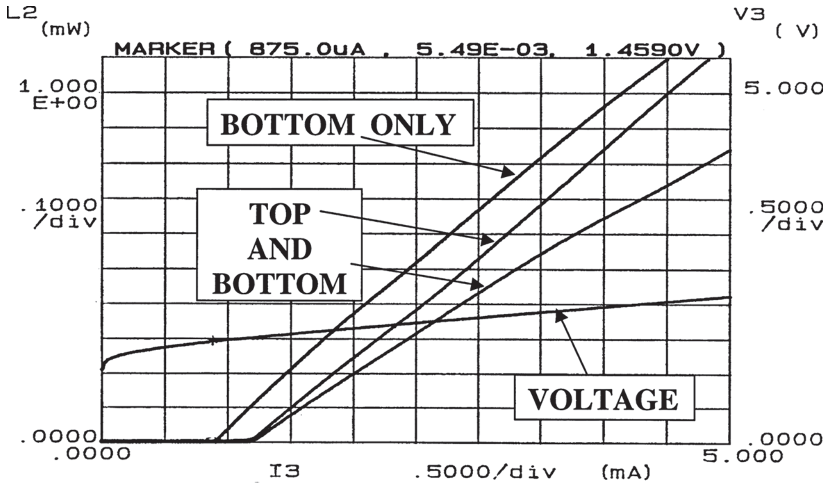


FIGURE 27.12 The *L–I* and *V–I* curves of a bottom-only and an adjacent bi-directionally emitting I^2-VCSEL (both 10 µm diameter aperture)

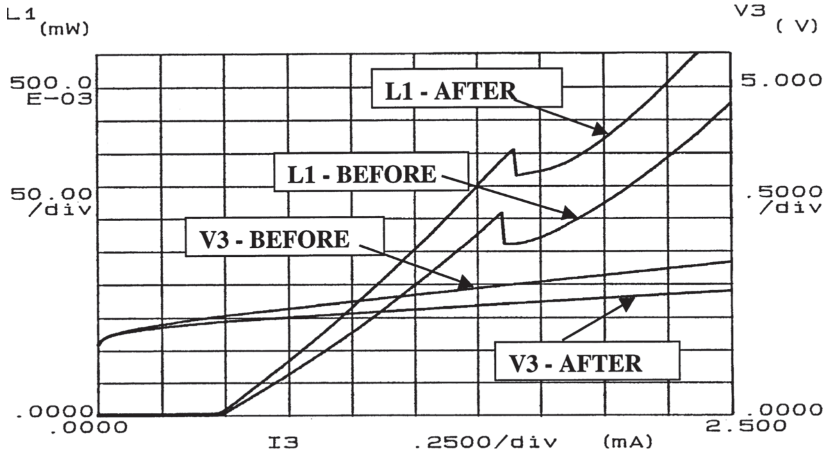


FIGURE 27.13 The characteristic of an I^2-VCSEL, obtained just before and then just after flip-chip bonding.

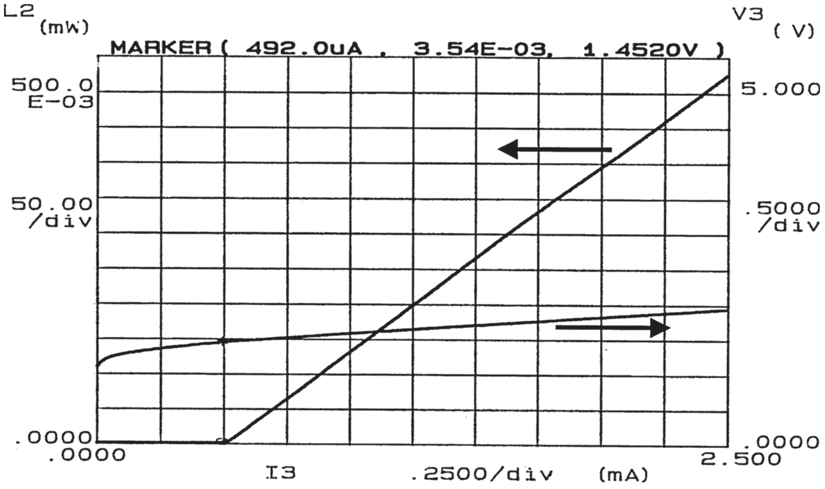


FIGURE 27.14 The *L–I* and *V–I* curves of a flip-chip bonded I²-VCSEL, with a proper AR-coating.

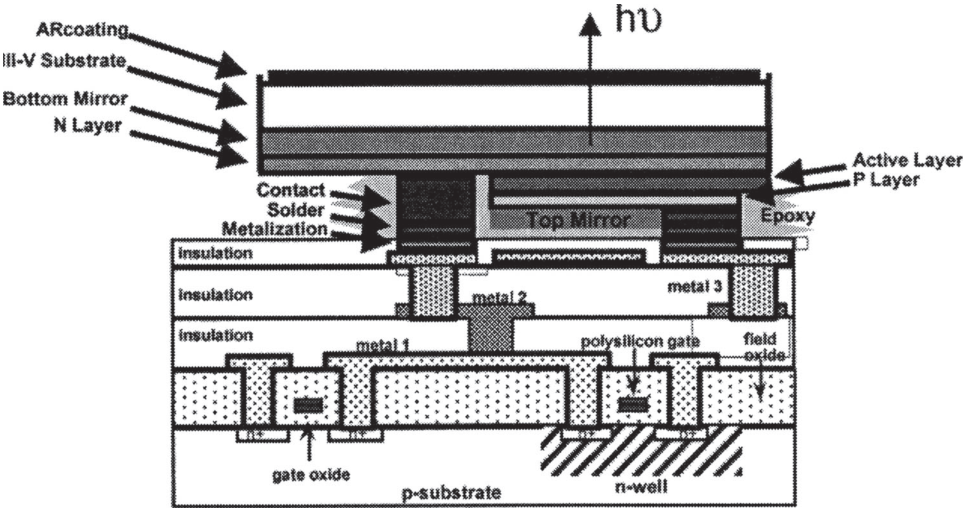

hυ
ARcoating
III-V Substrate
Bottom Mirror
N Layer
Active Layer
P Layer
Contact
Solder
Metalization
Top Mirror
Epoxy
insulation
insulation
insulation
metal 3
metal 2
metal 1
polysilicon gate
field oxide
gate oxide
p-substrate
n-well

FIGURE 27.15 Schematic diagram of a flip-chip I²-VCSEL bonded to CMOS. The I²-VCSEL is bonded over active gates.

the GaAs substrate. A schematic cross section of the assembled structure is depicted in Figure 27.15. The I²-VCSELs were bonded directly to the top-level metal on the CMOS chip and did not interfere with the operation or the layout of the underlying circuits [26]. A microphotograph of part of a 2 × 10 bottom emitting array prior to bonding is shown in Figure 27.16. A photograph of the assembled structure can be seen in Figure 27.17a. The CMOS drive circuits allowed each laser to be independently modulated and tested by external control. Figure 27.17b is a close-up photograph of the GaAs substrate face of a CMOS/VCSEL chip, with the entire array forward-biased, sub-threshold and with a single device biased over threshold and lasing. All the VCSELs were operational on some chips but, typically, one to two devices per chip did not lase due to observable processing errors. The operational VCSELs continued to lase through several weeks of testing, making infant mortality an unlikely issue. Most of the integrated I²-VCSEL arrays comprised 10 μm diameter devices, whose threshold

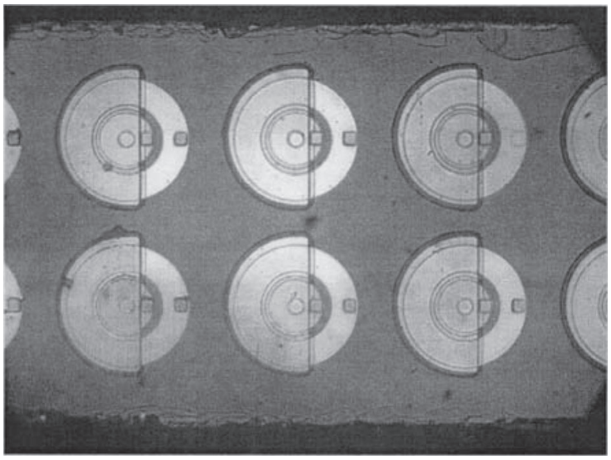
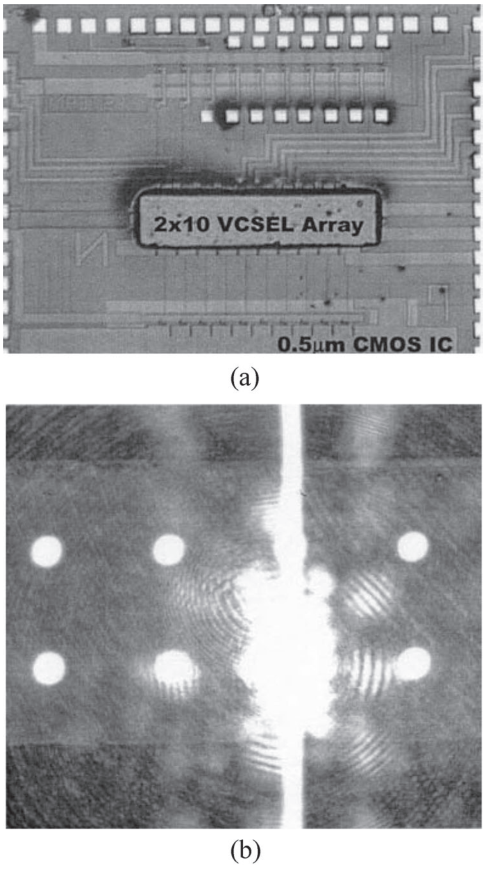


FIGURE 27.16 Microphotograph of a portion of a 2 × 10 I²-VCSEL array prior to flip-chip bonding.

voltages and currents, after bonding, remained at ~1.5 V and ~1 mA, respectively. Both of these values are well within the capabilities of commercial CMOS technologies. The optical outputs of these devices, with 10 μm diameter apertures, were multi-mode, similar to those of the top-emitters, described previously. At room temperature, their outputs typically exceeded 4 mW before rollover occurred at input currents between 16 and 18 mA. These data confirmed that good electrical and thermal contact to the lasers was made via the flip-chip bonds.

To investigate the temperature effects on a CMOS/VCSEL chip and its reliability, one such chip was placed directly on a thermo-electric cooler and light *versus* current (*L–I*) curves, as well as voltage *versus* current (V–I) curves were recorded for various temperatures. These appear in Figure 27.18. Note that at 80°C, the output is still nearly 1 mW. No appreciable change in performance developed after temperature cycling. The wavelength shift in the output of a flip-chip bonded VCSEL *versus* temperature and *versus* electrical input power was also measured. The resulting slopes for the linear fits to the two sets of data were 15°C nm^{-1} and 0.07 nm mW^{-1}. The product of these two values is the thermal resistance of the hybrid integrated I^2-VCSEL and was approximately 1°C mW^{-1}.

The VCSEL driver, on the CMOS/VCSEL chips, was a simple two-transistor version, based on a current-shunting principle, which provides a low-area, tunable-power circuit. A schematic illustration appears in Figure 27.19. The measured small-signal bandwidth of the driver was on the order of 2 GHz when implemented in a 0.5 μm CMOS process. In this circuit, a PMOS transistor sends an adjustable amount of current through the VCSEL and an NMOS transistor can then shunt the current into and out of the laser. Initially fabricated drivers were able to provide up to 6 mA of current to their corresponding lasers. For bit-rate measurements, a commercial 2 GHz detector/receiver monitored the output of each individual VCSEL output. The coupling between a laser and a receiver was through a 5 m long, 62.5 μm-core multi-mode fibre patch cord. The CMOS/VCSEL transmitter was driven by ($2^{23} - 1$) bit long, pseudo-random bit stream. The bit error rate was $<10^{-11}$, when the I^2-VCSEL was biased above threshold. Without the pre-bias, adjusted by a change in V_{tune} (Figure 27.19), the transmitter would still operate at 1.25 Gbit s^{-1}, but the eye pattern would then indicate a turn-on delay of ~180 ps. The total



(a)

(b)

FIGURE 27.17 (a) I^2-VCSEL array after bonding to CMOS chip; (b) microphotograph of portion of the top face of a CMOS/VCSEL chip, taken with one I^2-VCSEL biased above threshold and lasing and the rest of the laser array forward-biased below threshold and just luminescing.

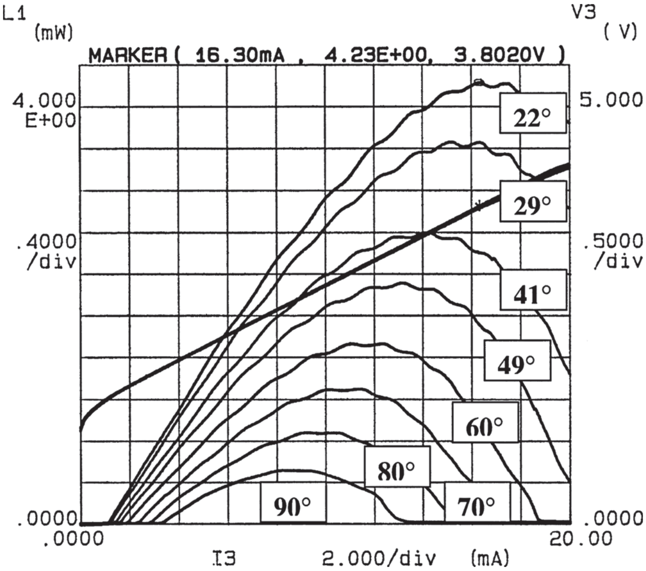


FIGURE 27.18 A set of *L–I* and *V–I* curves showing the temperature dependence of a flip-chip-bonded CMOS/VCSEL chip.

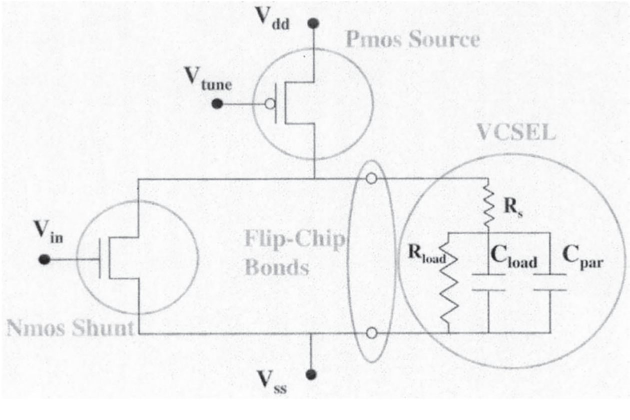


FIGURE 27.19 A circuit diagram of the CMOS/VCSEL transmitter. The equivalent circuit for the VCSEL includes a series resistance (R_S) and a parallel parasitic capacitance (C_{par}); both conspire to limit the modulation speed of the VCSEL.

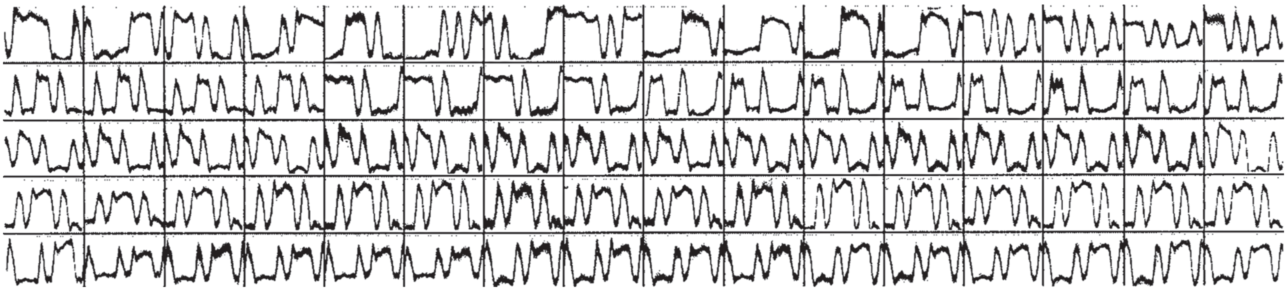


FIGURE 27.20 Data from a 16 × 5 array of VCSELs with individual CMOS drivers when simultaneously modulated by circuits at 1 Gbit s^{-1} with 32-bit pseudorandom bit patterns.

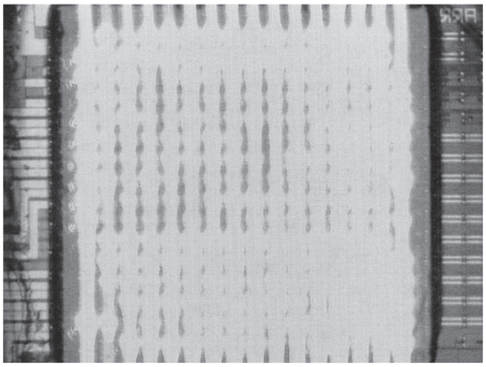


FIGURE 27.21 16 × 16 array of VCSELs under forward bias, with all but 10 of the VCSELs in the array lasing.

power dissipated per optical output, by the laser, the driver and all the on-chip buffers, was 17.5 mW, from a 3.3 V supply, for a −6.9 dBm optical power output at 1.25 Gbit s^{-1}.

These experiments validated the design methodology for the VCSELs and their integration to electronic circuits. Subsequently, much larger CMOS/VCSEL chips with arrays of 16 × 16 I^2-VCSELs of the same design were also demonstrated and reported [27]. Figure 27.20 shows data from an array of 80 VCSELs being modulated at 1 Gbit s^{-1} per laser, corresponding to an 80 Gbit s^{-1} aggregate throughput. The array was part of a larger 16 × 16 array of VCSELs that was integrated to a CMOS chip containing an array of drivers. Figure 27.21 shows the 16 × 16 array of 970 nm VCSELs simultaneously being forward-biased (above threshold) by a corresponding array of driver circuits on a 0.5 μm CMOS chip. The flip-chip-bonded VCSELs indicated a 3 dB bandwidth of 4.2 GHz, limited by the VCSEL device itself, and a crosstalk suppression of approximately −20 dB. The thermal impedance of the devices was approximately 1°C mW^{-1}. Parallel operation of 80 VCSELs at 1 Gbit s^{-1} per device was accomplished, proving that such lasers could be used for high bandwidth, parallel communication directly from a CMOS circuit.

27.4 Design Modifications for High-Speed P^{2}I^2-VCSELs

One potential issue with intra-cavity-contacted lateral current injection schemes is their tendency to increase the parasitic capacitance of the VCSELs (note the equivalent circuit for the VCSEL in Figure 27.19). This increases the *RC* time constant of the devices and, therefore, increases their rise- and fall-times, hence lowering modulation speed. In addition, it is more difficult to obtain a uniform current density distribution in the vertical aperture. This is especially true when the lateral current injection is radially symmetric. Holes travel essentially horizontally in their p-doped layer from the p-metal contact towards the aperture and electrons also travel essentially horizontally, along parallel radial paths, in their n-doped layer from the n-metal contact towards the aperture. The holes and electrons only flow in counter directions vertically inside the aperture region. In such situations, the current density can be greater at the rim of the aperture and lower at the centre, whereas, for VCSELs, it is generally preferable that the current density be uniform or even greater at the centre of the aperture. As explained later, both of these problems were addressed in a modified P^{2}I^2-VCSEL design wherein the conductivity of the bottom-doped layer (usually the n-layer) was locally suppressed and where the holes and the electrons were restricted to counter-flowing lateral paths (Figure 27.22).

In a typical VCSEL, the p-doped layer and its contact constitute one plate and the n-doped layer and its contact is the other plate of the built-in capacitor of the device. Options for reducing the device capacitance are limited. Since the lengthening of a VCSEL cavity is detrimental to the device performance, intra-cavity plates cannot be placed farther apart. The material between the plates is essentially fixed, so the dielectric constant between the plates is fixed. The area of each plate must be large enough to cover the active area and then must allow room for the metal contacts as well as additional area for fabrication tolerances. Therefore, the plate areas are several times as large as the active areas and, in radially symmetric designs, can be more than ten times as large and cannot be reduced without making the fabrication much more difficult. Nevertheless, all the excess area contributes to excess parasitic capacitance, except when the overlap of the two plates is reduced. In VCSEL designs, where the top mirror is deposited and not epitaxially grown, the reduction of the overlap area can be accomplished by a relatively shallow ion implantation that is patterned to locally suppress the conductivity of the bottom-doped layer (usually n-doped) under most of the top-doped layer (usually p-doped). The overlap area can be made only slightly larger than the active area. The reduction in capacitance can be as large as tenfold, assuming that a low series resistance (R_s in Figure 27.19) can be maintained.

When this implantation technique is employed, the VCSELs can have their p-contact to one side of the active area (the left-hand side, for instance) with the n-layer under it suppressed. The n-contact can then be placed on the other side (e.g. the right side) with the p-layer etched away. Under forward bias, holes will then travel from their contact towards the aperture mostly from left to right and the electrons will travel from their contact predominantly from right to left. If the conductivities of both of the doped layers were about the same, then all current paths through the aperture would be equivalent, the current would be carefully guided from one contact to the other and the current density would be more uniform than in the radially symmetric case [24].

27.5 Results from the First High-Speed P^2I^2-VCSELs

The previous processing step was added as an option to the I^2-VCSEL fabrication sequence and devices were made on the same wafer with a mask set containing both the old I^2-VCSELs and the modified high-speed P^2I^2-VCSELs. The new designs showed better cw characteristics and, as expected, were much faster [28]. Figure 27.23 shows the small signal bandwidth reaching 10.5 GHz, at a bias current of ~4.2 mA, for a new 6 µm aperture device, lasing single mode, at ~964 nm. During the experiment, the device, still part of a wafer, was just resting

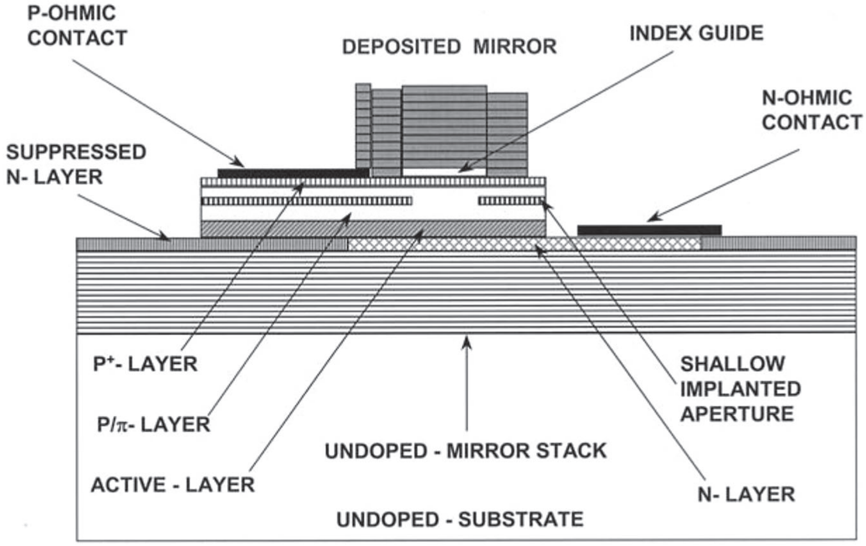


FIGURE 27.22 A schematic diagram of a cross section of the new high-speed P^2I^2-VCSEL design. The conductivity of the n-layer is suppressed under the p-contact, so as to reduce parasitic capacitance and the lateral current injection layers and their contacts are so configured as to make all current paths through the aperture essentially equivalent.

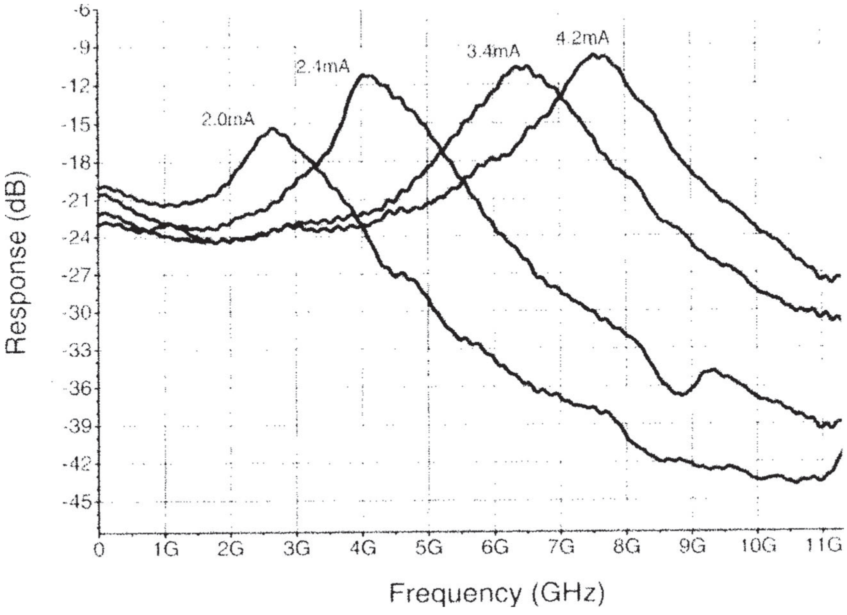


FIGURE 27.23 The small signal modulation response of a high-speed P^2I^2-VCSEL at various bias currents, showing a 3 dB bandwidth of ~4 GHz at 2 mA, ~6 GHz at 2.4 mA, ~9 GHz at 3.4 mA and ~10.5 GHz at 4.2 mA.

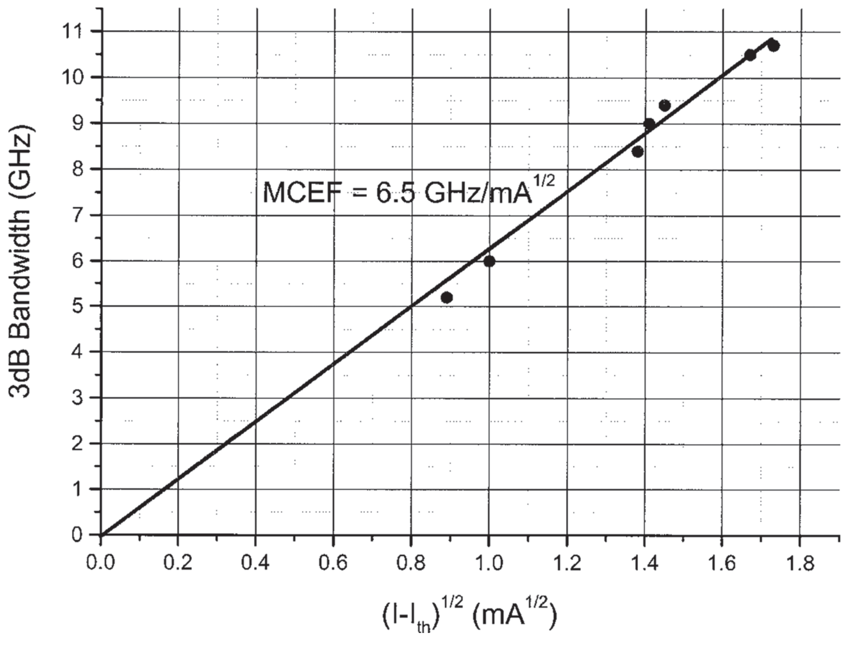


FIGURE 27.24 The MCEF for a 6 µm aperture, single-mode, high-speed P^{2}I^2-VCSEL (defined as the slope of the linear fit to the plot of 3 dB bandwidth *versus* $(I-I_{\text{th}})$).

on a probe-station chuck, with no special heat-sinking or cooling. Heating effects limited the useful range of the device to bias currents <4.5 mA. Eventually, a careful alignment of the gain peak with the Fabry–Pérot cavity resonance wavelength and the use of flip-chip bonding for better heat sinking should reduce these heating effects and thus extend the useful range of devices. The 3-dB bandwidths of the same device, for various currents throughout the useful range, are plotted *versus* the square root of $(I_{\text{bias}}-I_{\text{threshold}})$ in Figure 27.24. The data fit a straight line over the entire present range of the device. This fact strongly implies that increases in the bandwidth will follow the reductions in the heating effects. The slope of the line in Figure 27.24 is the modulation conversion efficiency factor (MCEF) for the VCSEL [29] and is approximately equal to 6.5 GHz $\sqrt{mA}$. The MCEF is a useful parameter for predicting the extent of possible improvements.

An HP lightwave signal analyzer with a built-in tracking generator, which allowed calibrated modulation response measurements from 0 to 18 GHz, collected the data shown in the previous two figures. The signal analyzer stimulated the VCSEL with a 40 GHz microwave probe through a bias-tee (>12 GHz), while a fibre probe collected the output light, sending it to a high-speed detector and amplifier (25 GHz), which then fed the signal to the analyzer. The response of the bias-tee, the photo-detector and the amplifier was recorded separately and was then subtracted from the data for the VCSEL. The expected small signal modulation characteristics of a laser can be approximated as

$$P(f)=\frac{F_{\mathrm{R}}^{4}}{\left((F_{\mathrm{R}}^{2}-f^{2})+\left(\frac{\Gamma}{2\pi}\right)^{2}f^{2}\right)\left(1+\frac{f^{2}}{f_{0}^{2}}\right)}$$

where F_{R} is the resonance (relaxation oscillation) frequency of the laser, Γ is the damping frequency and f_0 is the parasitic (3 dB) cut-off frequency, which includes not only the *RC* delay but also possible carrier-transport limitations [30]. A three-parameter (F_{R}, Γ and f_0) fit to each of the data curves in Figure 27.23 was performed, and the values for those parameters were extracted. The best fit for f_0, for all of the curves, was 20 GHz. This value was in general agreement with the estimate of the 3*RC* time constant for that VCSEL, which was ~60 ps, and supported the assertion that the redesign for high speed (as described in the previous section) was effective. To test the premise that heating effects limited the bandwidth by raising the device temperature, the previous experiment was repeated at an ambient temperature 10°C below room temperature by placing the chip on a controlled temperature stage, to partially counterbalance the heating effects. An increase in the maximum resonance frequency of approximately 2 GHz (to 9.5 GHz) was observed, in support of this hypothesis.

The extracted values for the resonance frequencies and the damping frequencies are plotted in Figure 27.25, as Γ *versus* F_{R}^2. The data points fit a line and the slope of that line is defined as the *K*-factor of the laser. The importance of the K-factor stems from the fact that it can be used to determine the maximum intrinsic 3 dB bandwidth of the laser, as follows [30].

$$f_{\max}=\frac{2\sqrt{2}\pi}{K}.$$

For the 6 µm aperture, high-speed P^{2}I^2-VCSEL under discussion, $f_{\max}$ = 95 GHz. This very large value should not immediately raise great expectations, since there are many effects that can limit laser bandwidths at values much lower than $f_{\max}$, including the already obvious heating effects and *RC* parasitics. Carrier transport limits, especially in quantum-well lasers, can also be important [31] and are usually indistinguishable

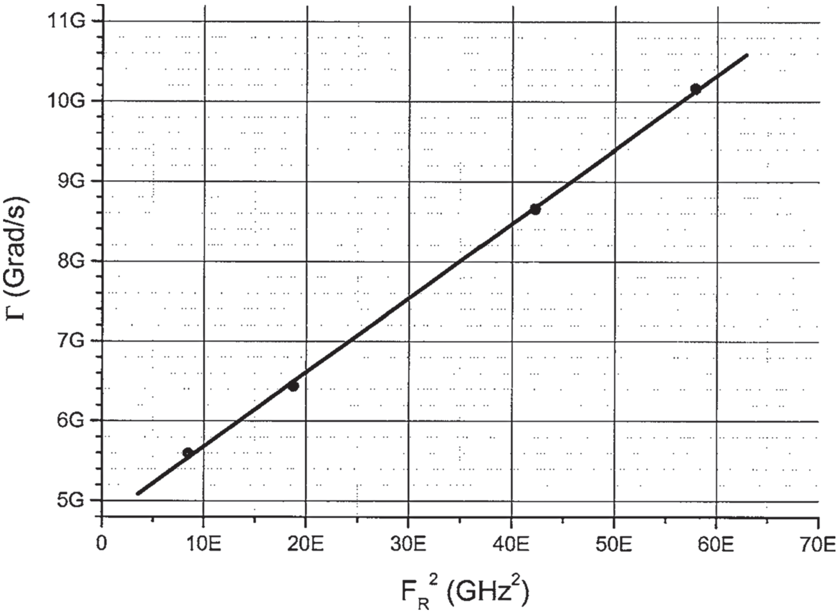


FIGURE 27.25 A plot of the damping rates *versus* the squares of the resonance frequencies, for a 6 μm aperture, single-mode, high-speed P^{2}I^2-VCSEL. These values come from the three-parameter fits to the small signal modulation responses of the device. The slope of the linear fit to the data is the *K*-factor of the laser, used to determine the maximum intrinsic 3 dB bandwidth of the laser, $f_{\max}$.

from *RC* delays. The large value of $f_{\max}$ does mean that significant bandwidth improvements may be possible and that, once the heating effects are eliminated, then f_0 (at 20 GHz) again becomes the limitation, either due to *RC* parasitics or carrier-transport problems.

27.6 Conclusions

This chapter has described the development of VCSEL-based optoelectronic-VLSI interconnects. This development is directly driven from increasing bandwidth requirements for next-generation telecommunications and data communications equipment, specifically high-end routers, optical cross-connect systems and large data servers, whose capacity and scalability are continually being challenged by the physical limitations of the inter-connects and switching sub-systems being employed in their boards, backplanes and chassis. VCSEL-based parallel optical links have already penetrated the market as cost-effective means to transport data inside such systems. This chapter has described a project whose purpose was to develop a fabrication platform that would reproducibly make high-efficiency VCSELs that easily integrate and operate with electronic circuitry and could, thus, provide fibre inter-connections directly to VLSI chips. To that end, a new design strategy was conceived, whereby VCSELs were processed with photolithographic precision (for reproducibility) and contained an intra-cavity contacted current guide (for easier integration) and an independent optical beam guide. The beam guide is a feature that can be made and optimized, independently of the current guide and the mirrors. The use of a patterned, deposited dielectric top mirror makes the design strategy feasible, by allowing the formation of crucial device features to occur near the surface with high precision and accuracy, before the top mirror is placed. Note that both the contacts, the dual lateral current injection layers and the requisite current aperture, are all intra-cavity-located. This helps the light to resonate between the mirror stacks as efficiently as possible as it passes back and forth through the current aperture, the contact openings and the many reflecting interfaces. The formation of the two guides is greatly facilitated when the top mirror is a dielectric stack, deposited and patterned at the end of the fabrication sequence.

Such P^{2}I^2-VCSELs have been fabricated in top, bottom and bi-directionally emitting configurations, in the wavelength range around 980 nm. Devices, with 6–10 μm diameter apertures, had ~0.4 to ~1.0 mA threshold currents, ~1.5 V threshold voltages and optical power outputs of 1 mW at ~2.5 mA input. The spectra of the lasers with ~10 μm apertures were rich in multiple transverse modes, but devices with ~6 μm apertures and below were single mode. Subsequently, bottom-emitting I^2-VCSELs were made and arrays (of 2 × 10 devices and then of 16 × 16 devices) were hybrid integrated, via flip-chip bonding, to CMOS chips with custom VCSEL drivers, in a 3.3 V, 0.5 μm CMOS technology. These CMOS/VCSEL chips had optical outputs per channel, at ~970 nm, of up to 4 mW at room temperature and up to 1 mW at 80°C, if driven externally, and about half as much if driven by the CMOS transmitter circuits which could deliver a maximum of 6 mA to the VCSELs. The transmitter drivers were able to modulate the VCSELs at data rates of 1.25 Gbit s^{-1} channel^{-1} with a bit error rate less than 10^{-11}. In one instance, 80 channels operated simultaneously at over 1 Gbit s^{-1} each.

The first versions of the I^2-VCSELs had a modulation speed limitation characteristic of intra-cavity-contacted VCSELs, attributable to high capacitance and insufficiently uniform current density distributions. The design was modified in subsequent versions to reduce capacitance by suppression of the n-layer conductivity under the p-ohmic contact region and by placing restricted area contacts on opposite sides of the active region, which improved the current distribution by making all current paths essentially equivalent. The initial results for the new version were a small-signal 3 dB bandwidth of 10.5 GHz, a parasitic cut-off frequency raised to 20 GHz and other characteristics strongly suggesting the potential for even greater improvements (such as an extracted maximum intrinsic 3 dB bandwidth, $f_{\max}$, of 95 GHz). More work is needed, however, to guarantee that such a VCSEL fabrication platform can indeed supply large volumes of reproducibly high yield and reliable VCSEL arrays that will ultimately allow low-power, high-bandwidth systems based on fibre and surface-normal optical inter-connects for data transport and silicon VLSI for switching.

Acknowledgments

Numerous colleagues at Bell Laboratories have contributed to this work. We especially appreciate the contributions of L. A. D'Asaro, J. E. Cunningham, W. Y. Jan, G. Giaretta, S. P. Hui, R. E. Leibenguth, B. J. Tseng, J. M. Vandenberg, J. D. Wynn and G. J. Zydzik, and the guidance of W. T. Tsang, W. H. Knox and M. C. Nuss.

REFERENCES

1. Tooley F. A. P. 1993 Challenges in optically interconnecting electronics *IEEE J. Selected Topics Quantum Electron.* **2** 3–13.
2. Ahadian J. F. and Fonstad C. G. Jr 1998 Epitaxy-on-electronics technology for monolithic optoelectronic integration: foundations, development, and status *Opt. Eng.* **37** 3161–74.
3. Pedrotti K. D. Beccue S., Haber W. J., Brar B. P., Robinson G. and Kilcoyne M. K. 1990 16 × 16 optical-fiber crossbar switch operating at 0.85 micron *J. Lightwave Technol.* **8** 1330–42.
4. Cinato P. and Young K. C. Jr 1990 Optical interconnects within multichip modules *Opt. Eng.* **32** 852–60.
5. Banwell T. C., Lehmen A. C. and Cordell R. R. 1993 VCSE lasers for parallel optical data links *IEEE J. Quantum Electron.* **29** 635–44.
6. Zeeb E., Moeller B., Reiner G., Ries M., Hackbarth T. and Ebeling K. J. 1995 Planar proton implanted VCSELs and fiber-coupled 2-D VCSEL arrays *IEEE J. Selected Topics Quantum Electron.* **1** 616–23.
7. Hartman D. H. 1986 Digital high-speed interconnects: a study of the optical alternative *Opt. Eng.* **25** 1086–102.
8. Desmulliez M. and Wherrett B. S. 1999 *Optical Interconnections and Parallel Processing: Trends at the Interface* ed E. P. Berthome and A. Ferreira (Dordrecht: Kluwer Academic) ch 2, pp. 19–48.
9. Miller D. A. B., Chemla D. S., Damen T. C., Gossard A. C., Wiegmann W., Wood T. H. and Burrus C. A. 1984 Novel hybrid optically bistable switch: the quantum well self-electro-optic effect device *Appl. Phys. Lett.* **45** 13–5.
10. Boyd G. D., Miller D. A. B., Chemla D. S., McCall S. L., Gossard A. C. and English J. H. 1987 Multiple quantum well reflection modulator *Appl. Phys. Lett.* **50** 1119–21.
11. D'Asaro L. A., Chirovsky L. M. F., Laskowski E. J., Pei S. S., Woodward T. K., Lentine A. L., Leibenguth R. E., Focht M. W., Freund J. M., Guth G. G. and Smith L. E. 1993 Batch fabrication and operation of GaAs-Al_xGa_{1-x} As field-effect transistor self-electro-optic effect device (FET-SEED) smart pixel arrays *IEEE J. Quantum Electron.* **29** 670–7.
12. Goossen K. W., Walker J. A., D'Asaro L. A., Hui S. P., Tseng B., Leibenguth R., Kossives D., Bacon D. D., Dahringer D., Chirovsky L. M. F., Lentine A. L. and Miller D. A. B. 1995 GaAs MQW modulators integrated with silicon CMOS *IEEE Photon. Technol. Lett.* **7** 360–2.
13. Kuo J. M., Wang Y. C., Goossen K. W., Chirovsky L. M. F., Hui S. P., Tseng B. T., Walker J., Lentine A. L., Leibenguth R. E., Livescu G., Jan W. Y., Cunningham J. E., D'Asaro L. A., Ron A., Dahringer D., Kossives D., Bacon D. D., Morrison R. L., Novotny R. A., Buchholz D. B. and Mayo W. E. 1997 Large array of GaAs modulators and detectors flip-chip solder bonded to silicon CMOS using InGaP as the selective etch stop for GaAs substrate removal *J. Crystal Growth* **175/176** 971–6.
14. Krishnamoorthy A. V. and Goossen K. W. 1997 Progress in optoelectronic—VLSI smart pixel technology based on GaAs/AlGaAs MQW modulators *Int. J. Optoelectron.* **11** 181–98.
15. Krishnamoorthy A. V. and Goossen K. W. 1998 Optoelectronic—VLSI: Photonics integrated with VLSI circuits *IEEE J. Selected Topics Quantum Electron.* **4** 899–912.
16. Nakahara T., Matsuo S., Fukushima S. and Kurokawa T. 1996 Performance comparison between multiple quantum well modulator based and vertical cavity surface emitting laser based smart pixels *Appl. Opt.* **35** 860–71.
17. Huffaker D. L., Deppe D. G., Kumar K. and Rogers T. J. 1994 Native-oxide defined ring contact for low threshold vertical-cavity lasers *Appl. Phys. Lett.* **65** 97–9.
18. Choquette K. D. and Hou H. Q. 1997 Vertical-cavity surface emitting lasers: moving from research to manufacturing *Proc. IEEE* **85** 1730–9.
19. Choquette K. D., Geib K. M., Ashby C. I. H., Twesten R. D., Blum O., Hou H. Q., Follstaedt D. M., Hammons B. E., Mathes D. and Hull R. 1997 Advances in selective wet oxidation of AlGaAs alloys *IEEE J. Selected Topics Quantum Electron.* **3** 916–25.
20. Morgan R. A., Lehman J. A. and Hibbs-Brenner M. K. 1996 Vertical-cavity surface-emitting lasers come of age *Proc. SPIE (San Jose, CA)* **2683** 18–29.
21. MacDougal M. H., Geske J., Lin C. K., Bond A. E. and Dapkus P. D. 1998 Low resistance intracavity-contacted oxide-aperture VCSELs *IEEE Photon. Technol. Lett.* **10** 9–11.
22. Lei C., Rogers T. J., Deppe D. G. and Streetman B. G. 1991 ZnSe/CaF_2 quarter-wave Bragg reflector for the vertical-cavity surface-emitting laser *J. Appl. Phys.* **69** 7430–4.
23. Chirovsky L. M. F., Hobson W. S., Leibenguth R. E., Hui S. P., Lopata J., Zydzik G. J., Giaretta G., Goossen K. W., Wynn J. D., Krishnamoorthy A. V., Tseng B. J., Vandenberg J. M. and D'Asaro L. A. 1999 Implant-apertured and index-guided vertical-cavity surface-emitting lasers (I^2-VCSELs) *IEEE Photon. Technol. Lett.* **11** 500–2.
24. Chirovsky L. M. F., Krishnamoorthy A. V., Hobson W. S., Lopata J., Goossen K. W. and D'Asaro L. A. 2000 Vertical-cavity surface-emitting lasers specifically designed for integration with electronic circuits *SPIE Critical Review on Heterogeneous Integration* ed E Towe, ch 2.

25. Krishnamoorthy A. V., Chirovsky L. M. F., Hobson W. S., Leibenguth R. E., Hui S. P., Zydzik G. J., Goossen K. W., Wynn J. D., Tseng B. J., Lopata J., Walker J. A., Cunningham J. E. and D'Asaro L. A. 1999 Vertical-cavity surface-emitting lasers flip-chip bonded to gigabits-per-second CMOS circuits *IEEE Photon. Technol. Lett.* **11** 128–30.
26. Krishnamoorthy A. V., Lentine A. L., Goossen K. W., Walker K. W., Woodward J. A., Ford J. E., Aplin G. F., D'Asaro L. A., Hui S. P., Tseng B., Leibenguth R., Kossives D., Dahringer D., Chirovsky L. M. and Miller D. A. B. 1995 3-D integration of MQW modulators over active sub-micron transmitter circuit *IEEE Photon. Technol. Lett.* **7** 1288–90.
27. Krishnamoorthy A. V., Goossen K. W., Chirovsky L. M. F., Rozier R. G., Chandramani P., Hobson W. S., Hui S. P., Lopata J., Walker J. A. and D'Asaro L. A. 2000 16 × 16 VCSEL array flip-chip bonded to CMOS VLSI circuit *Photon. Technol. Lett.* **12** 1073–5.
28. Krishnamoorthy A. V., Chirovsky L. M. F., Hobson W. S., Lopata J., Shah J., Rozier R., Cunningham J. E. and D'Asaro L. A. 2000 Small-signal characteristics of bottom-emitting intracavity contacted VCSELs *Photon. Technol. Lett.* **12** 609–11.
29. Chen T. R., Zhao B., Eng L., Zhuang Y. H., O'Brien J. and Yariv A. 1993 Very high modulation efficiency of ultralow threshold current single quantum well lasers *Electron. Lett.* **29** 1525–6.
30. Olshansky R., Hill P., Lanzisera V. and Powazinik W. 1987 Frequency response of 1.3 μm InGaAsP high-speed semiconductor lasers *IEEE J. Quantum Electron.* **23** 1410–8.
31. Nagarajan R., Ishikawa M., Fukushima T., Geels R. S. and Bowers J. E. 1998 Epitaxy-on-optoelectronics technology for monolithic optoelectronic integration: foundations, development, and status *Opt. Eng.* **37** 3161–74.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

28

Advances in Laser Satellite Communications

Hamid Hemmati

CONTENTS

28.1 Introduction 383
28.1.1 Lasercom Benefits 384
28.1.2 Lasercom Challenges 384
28.1.3 Technology Maturity 385
28.1.4 Link Budget 385
28.2 Technology and Design Drivers 386
28.2.1 Laser Beam Spatial Acquisition, Tracking, Pointing and Stabilization 386
28.2.2 Rugged Opto-Mechanical and Thermo-Mechanical Optics 388
28.2.3 Signal Detection 389
28.2.4 Modulation and Coding 390
28.2.5 Flight Laser Transmitter 391
28.2.6 The Atmospheric Channel 391
28.2.6.1 Mitigating Effects of Turbulence/Scintillation 392
28.3 Ground Station 392
28.3.1 Ground Receive Telescope 392
28.3.2 Ground Uplink Transmitter 392
28.3.3 Ground PAT 393
28.4 Applications 393
References 394

28.1 Introduction

Globally, the Internet traffic supported by wireline and wireless communications exceeds multi-zettabyte (8×10^{12} bits) of data level [1]. By 2020, this figure is projected to reach nearly two orders of magnitude greater than the global Internet traffic in 2005. To date, the fibre-optics communications technology governs high-data-rate access and backbone networks with tens of Tbps capacity. However, scaling of fibre-optic network to the rural areas of the globe is either impractical or has not been implemented yet due to excessive cost or geological obstacles.

Satellite communication offers the potential to bridge this gap and particularly service backhaul, and rural areas. The demand from Earth-orbiting satellites for delivery of broadband and access links as well as from the Earth-observing instruments in space is already in place and growing exponentially. Conventional satellites and high-altitude-platforms use radio-frequency (RF) and microwave (MW) frequencies for communications, and primarily from the geo-synchronous-orbit (GEO). Despite high reliability, current satellite communications capacity is a small fraction of what fibre-optic links can now provide.

At the time of this writing, the ViaSat-2 satellite with a total capacity of 250 Gbps provides the highest capacity based on use of a few hundred beams, with a data rate of 1–2 Gbps per beam [2]. By 2022, ViSat-3's launch is planned with a total capacity (uplink plus downlink) of greater than 1 Tbps. This substantial uplink and download link capacities are made possible via hundreds of low-data-rate beams [3]. Also, in this time frame, constellation(s) of LEO satellites are promising multi-Tbps aggregate capacity [4].

Due to the explosive growth in the level of information that will be communicated globally, future spacecraft will require radical changes in the way that data are moved around. RF and MW satellite telecom systems technologies are currently at the limits of their state-of-the-art and are at the same time hampered by the available spectrum limitations and regulations. The laser communications (lasercom) technology, with multi-THz of bandwidth available for telecom, and a spectrum that is currently unregulated, provides an opportunity to augment the conventional RF/microwave communications technology with a substantial capacity [5,6]. With the lasercom technology there are also major opportunities in reducing the aperture size, weight and power. As noted later in this chapter, since the year 2000, a number of successful demonstrations from the Earth orbit and from the Moon have proven that this technology is ready for operational use.

In near-Earth telecommunications, in applications, such as broadband Internet delivery that requires high (≥99.9%)

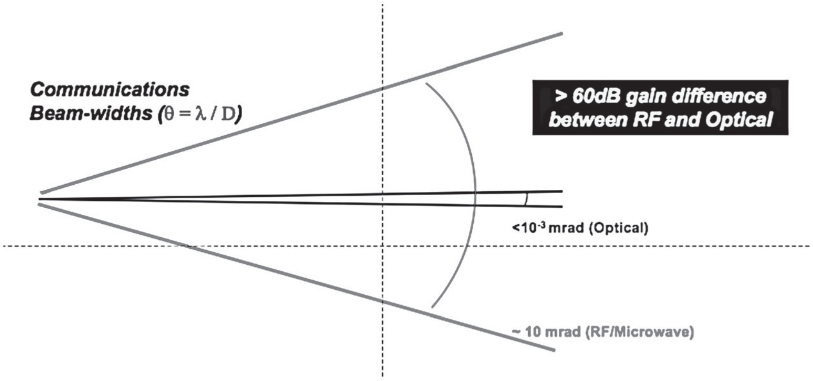


FIGURE 28.1 Not to scale. Nearly three orders of magnitude difference in beam divergence between RF and optical.

link availability, the use of lasercom downlink and uplink to the satellite is not practical. But, it would function perfectly well for lower link availability applications, such as the earth-observing or other sensor-based satellites. Gateway links to GEO satellites at rates of hundreds of Gbps to multi-Tbps from spatially diverse ground stations followed by RF/microwave downlink could appreciably enhance the current capacity limits of GEO satellites. Similarly, satellite cross links at the rate of tens to hundreds of Gbps can enable satellite constellations with a total capacity exceeding 10 Tbps [7].

This chapter summarizes various sub-system technologies that constitute a lasercom system and describes the basic tenets, state-of-practice as well as its technology trends.

28.1.1 Lasercom Benefits

Low-beam divergence. A major contributor to the advantage of laser communications with 200–300 THz frequency over the shorter frequencies (1–100 GHz) stems from nearly over three orders of magnitude smaller Optical beam divergence as determined by reduced diffraction of the beam. Consequently, the space loss factor in the link budget is reduced appreciably, as shown in Figure 28.1. Major benefits are reduced aperture (antenna/telescope) diameter, mass and power for the flight transceiver. The far-field beam divergence (in radians) is given by $\theta = \lambda / D_T$, where λ is the laser wavelength and D_T is the diameter of the transmit aperture that is filled by the beam.

Availability of bandwidth. A combination of hundreds of THz bandwidth and the wavelength division multiplexing (WDM) technology provides ample bandwidth at the optical spectrum, compared with highly limited <1 GHz to about 10 GHz for the different RF, MW and millimetre-wave (S, X, Ku, Ka, Q, V, E and W) bands. The enormous bandwidth of optical spectrum can be used to deliver Tbps data rates and allows for the use of bandwidth-intensive modulation and coding schemes that would not be possible otherwise.

Lack of spectrum regulation. Severe regulations on the use of the spectrum have long hindered the widespread use of RF and MW frequencies. A major benefit of the optical frequencies is that both of these obstacles are alleviated as greater than 100 THz of spectrum is available for use. Besides eye-safety regulations for some applications, there are currently no regulations on the use of the spectrum.

Use of commercially available fibre-optics components. Laser communications provide the ability to launch the received laser beam into a fibre and take advantage of the significant level of components developed for the fibre optics industry. Due to the exponential global growth of the Internet, the fibre optics technology is expected to continue its evolution.

Low probability of intercept and jamming. Tactical applications can benefit since lasercom beams are difficult to intercept and jam relative to RF and MW systems.

28.1.2 Lasercom Challenges

- Atmosphere affects the propagating laser beam in two ways: attenuation (e.g. clouds or fog) and the index of refraction-fluctuation affects (e.g. turbulence and scintillation), resulting in deep long-term or short-term signal fades, and beam break-up, among examples. Other than station diversity, there is currently no mitigation technology for attenuation, but a number of technologies have been devised to effectively mitigate downlink beam scintillation effects. Inter-satellite links operating through the vacuum environment are not affected by atmospheric effects.
- Due to the more energetic nature of the optical photons, optical receivers are quantum noise limited compared with the RF/MW frequencies that are thermal noise limited. As a result, RF receivers are more power-efficient than optical receivers (e.g. detecting $\sim 10^{-21}$ W vs. $\sim 5 \times 10^{-19}$ W for optical). Therefore, an optical communications system with beam of the same divergence as an RF communications beam will perform poorly. The narrow beam divergence of laser beams provides its main advantage. Energetic optical photons, however, may be taken advantage of in the form of noiseless single-photon detection, which is of interest for communication scenarios involving photon-starved regimes.
- Except for short distances (tens of metres) line-of-sight is required to establish the link.
- With a few exceptions, off-the-shelf components qualified for space use are not available.
- Optical communications ground station infrastructure is limited but growing slowly.

TABLE 28.1

Examples of Successfully Performed Spacecraft Lasercom Links

Project Name	LCE	SILEX	LUCE	1st Gen. LCT	2nd Gen LCT	LLCD
Host Spacecraft	ETS-VI	SPOT-4 ARTIMIS	OICETS (to Ground)	TerraSARX / NFIRE	Alphasat	LADEE
Demonstration Year	1995	2001	2006	2008	2013	2013
Orbits(s)	Near-GEO	LEO/GEO	LEO	LEO–LEO	LEO-GEO	Lunar
Downlink/cross-link Data rate (Mbps)	1	2 (GEO)	50	5625	1800	620
Link range (1000 km)	~36	36	0.6	5.1	~36	360–400
Downlink power (mW)	14	60	100	700	2200	500
Uplink rate (Mbps)	1	50 (LEO)	~2	5,625	1,800	20
Detection technique	Direct-Detection	Direct Detection	Direct Detection	Coherent Homodyne	Coherent Homodyne	Amplitude
Flt terminal mass (kg)	22	157	140	35	56	~30
Power consumption (W)	81	150	220	120	185	50–140
References:	[10]	[11]	[12]	[13]	[14]	[15]

Regardless of the challenges, which also exist in different forms at other telecom frequencies, lasercom technology offers a powerful telecom technology that can deliver tremendous volumes of data across the globe.

28.1.3 Technology Maturity

For decades, lasercom was an emerging technology maturing slowly with each space demonstration. Today, laser communication from the earth-orbit and between different earth-orbits is considered a mature technology to the point that such systems are available commercially and are deployed for operational use [8]. Certainly, further technology maturation will continue to occur as new or improved components and processes become available towards simplifying the system and reducing the development cost, size, weight and power consumption. Automated integration and test would greatly help in volume manufacturing of the flight transceiver units.

Since the early days of laser development, both terrestrial and space lasercom systems have been under development in the United States, Europe and Japan [9]. Only in the past two-decades major lasercom systems from Earth-orbiting spacecraft were demonstrated successfully, as outlined in Table 28.1.

28.1.4 Link Budget

As any other communications system, the Optical link budget is an important step in systems engineering process of designing a lasercom link that meets the requirements with comfortable margin. The communications portion of the link consists of a transmit channel, the optical channel between the transmitter and receiver, and the receive channel.

Transmit Channel. The data to be transmitted are impinged on the laser transmit beam via modulation of the laser. Today's forward-error-correction (FEC) codes are within 1 dB of the Shannon limit [16] with coherent receivers and those using photon-counting (single-photon sensitive) detection. Link overhead is involved in implementing the codes at high communication rates. The coherent homodyne phase-shift-keying (PSK) demonstrates the best photon efficiency with approximately 3 bits-per-photon sensitivity [17]. Laser's modulation extinction ratio (MER), the amplifier's amplified spontaneous emission and non-linear spectral broadening are among the factors that could limit the efficiency of the transmit channel.

Optical Channel. Clear-line-of-sight and cloud-free line-of-sight are requirements of any long-range optical communications system. Two major optical channel phenomena, namely attenuation and scintillation, may introduce major signal attenuation or temporary fades in the communications link. Attenuation may come from clouds, fog, heavy rain and snow, and any atomic or molecular absorption within the atmosphere [18]. At this time, there is no effective mitigation technique against attenuation through absorption and scattering. Scintillation/turbulence can cause transient fades with magnitude greater than 20 dB. The net effect (as explained in more detail later in this chapter) includes beam break-up, beam-dancing at the focal plane of the receiver, and highly inefficient coupling of the received beam into a single-mode receiver (e.g. a single-mode fibre) [19]. Multi-beam transmission and the adaptive-optics technology are among the effective mitigation techniques to counter the scintillation phenomenon.

Receive Channel. As inferred from Table 28.2, inductance and capacitance of the detection element cause the detector size to decrease substantially as the data rate increases. The smaller diameter of the photodetector can result in inefficient collection of the received signal that has propagated through the atmosphere. Detector's detection efficiency, various noise parameters inherent in the detector and the detector timing

TABLE 28.2

Examples of Nominal (but Not the Best Achievable) Detector Sensitivities for PIN and APD (Avalanche Photo-Diode) Detectors

	Sensitivity (dBm)		
Data Rate (Gbps)	PIN	APD	Detector Diameter (mm)
0.155	–41	–48	300–500
1.25	–32	–38	75–200
2.5	–29	–27	30–50
10	–20	–27	30–50

jitter which limits the data rate are among the limiting factors in the receiver.

Not including various loss terms in the transmit, the optical, and the receive channels, power received by the detector (P_{R}) as a function of the transmit power (P_{T}), diameters of the transmit and receive apertures (A_{T} and A_{R}), the transmitter's wavelength (λ_{Υ}) and the link range (R) are given by [20]

$$P_{\mathrm{R}} \approx \left(P_{\mathrm{T}} A_{\mathrm{T}} A_{\mathrm{R}}\right) / (\lambda_{\mathrm{T}}^2 R^2)$$

In the logarithmic domain, the communication link budget is simply a subtraction of the sum of all loss terms in different channels from the sum of all gain terms (e.g. laser transmit power, the gain of the transmit and receive antennas, and signal recovered via FEC). Given a set of system parameters (e.g. antenna gains) and the desired link margin, the required power at the transmitter can be derived from the link's signal-to-noise ratio (SNR) and the required bit-error-rate (BER) [21]. For an asymmetric link, separate link budgets are required for each link direction. Similarly, a dedicated budget analysis is required for the pointing, acquisition and tracking (PAT) channel to ensure that all loss factors are taken into account and adequate margin is allocated for the PAT process. Link budget drivers include the following:

1. set by the customer's requirements and specifications, including the link range, link data rate and BER, mass, power consumption and volume (restricting laser output power and antenna diameters);
2. set by physics, for example, beam divergence and the associated space loss, which is also driven by the link distance;.
3. flexibility (knobs) of the system design, for example, aperture diameters at each end of the link, and the transmit laser power and
4. hardware quality variables, for example, optics wavefront quality (Strehl ratio), laser's MER, photodetector's sensitivity and detectivity, and multiple atmospheric loss terms that may be present for that specific link.

Including the various loss terms, the received signal power (P_{R}) can be derived from the following:

$$P_{\mathrm{R}} \approx \left(P_{\mathrm{T}} G_{\mathrm{T}}\right) \cdot \left(L_{\mathrm{T}} L_{\mathrm{D}} L_{\mathrm{A}} L_{\mathrm{F}} L_{\mathrm{AO}} L_{\mathrm{P}} L_{\mathrm{S}} L_{\mathrm{TR}} L_{\mathrm{SL}} L_{\mathrm{R}} L_{\mathrm{I}}\right) \cdot \left(G_{\mathrm{R}}\right)$$

where P_{T} and G_{T} are the transmitter's laser power and gain, respectively. G_{T} is given by:

$$G_{\mathrm{T}} = 2/\alpha^2[\exp(\alpha^2) - \exp(\alpha^2\gamma^2)]^2 \cdot (\pi D_{\mathrm{T}}/\lambda)^2$$

where α is the optimal truncation ration of the transmit optics and $\gamma = d/D$ is the ratio of the secondary mirror diameter to that of the primary mirror. L_{T} is the transmitter losses, L_{D} is the distance loss, L_{A} represents the atmospheric losses, L_{F} stands for fade losses due to atmospheric scintillation, L_{P} stands for the pointing losses, L_{AO} represents losses incurred by the introduction of adaptive optics (AO) at the receiver, L_{S} is for the optics quality (Strehl ratio) losses, a measure of the losses introduced by the wavefront aberration and given by the dimensionless parameter $S = \exp\left[-(2\pi\sigma/\lambda_{\mathrm{T}})^2\right]$, where σ is the root-mean-square (rms) path difference for an aberrated optical element with respect to an ideal optical design, L_{TR} represents losses incurred from tracking error, L_{SL} is the space loss stemming from antenna (lens or telescope) diameter's diffraction given by $(\lambda_{\mathrm{T}}/4\pi R)^2$, L_{R} is for the receiver losses, and L_{I} stands for various system implementation losses. G_{R} is receiver gain given by $G_{\mathrm{R}} = (1-\gamma^2) \cdot (\pi D_{\mathrm{R}}/\lambda)^2$.

28.2 Technology and Design Drivers

Figure 28.2 shows a high-level free-space optical communications signal flow block diagram from a satellite to ground and vice versa. The flight terminal (transceiver) interfaces with the host satellite platform through various mechanical, structural, electrical and thermal contacts that in turn may impart significant shock and vibration onto the lasercom terminal.

Depending on the location of the spacecraft in space, both the spacecraft and the lasercom transceiver may experience high degree of radiation from energetic protons and electrons to an assortment of other cosmic background radiations that may be present in that environment. These could be highly damaging to the transceiver and the spacecraft, if not tested and made to work in that particular environment, and at the particular doses of radiation that could also be accumulative in certain electrical and optical components. The optical channel in between the spacecraft and the earth surface consists of vacuum and the turbid atmosphere. Both turbid and scintillating atmosphere within the optical channel can cause severe impairments of the propagating laser beam to and from the ground station. Performance degradation at either ends of the link and due to any component in either transceiver may directly result in degradation of the link and adversely affect the link performance.

28.2.1 Laser Beam Spatial Acquisition, Tracking, Pointing and Stabilization

Precise laser beam pointing, on the order of a few microradians, stemming from laser's narrow beamwidth is a major design driver for laser communications transceivers. Host platform's jitter and positional degrees of freedom have to be precisely compensated while maintaining a robust opto-mechanical and thermos-mechanical optics assembly that mitigates even minute amounts of undesired motion. Due to the cross-velocity between the satellite and ground or other receiving stations, generally a compensating point-ahead (or pint-behind) angle has to be introduced along with laser beam-pointing. The magnitude of the point-ahead-angle can be several orders of magnitude larger than the lasercom beamwidth. The requirement for highly precise laser beam-pointing is arguably the most challenging aspect of a lasercom link with major design-driver consequences.

Low mass and power consumption, high transmitter and receiver efficiency, and sufficient lifetime of active elements in the space environment are among the other technology drivers.

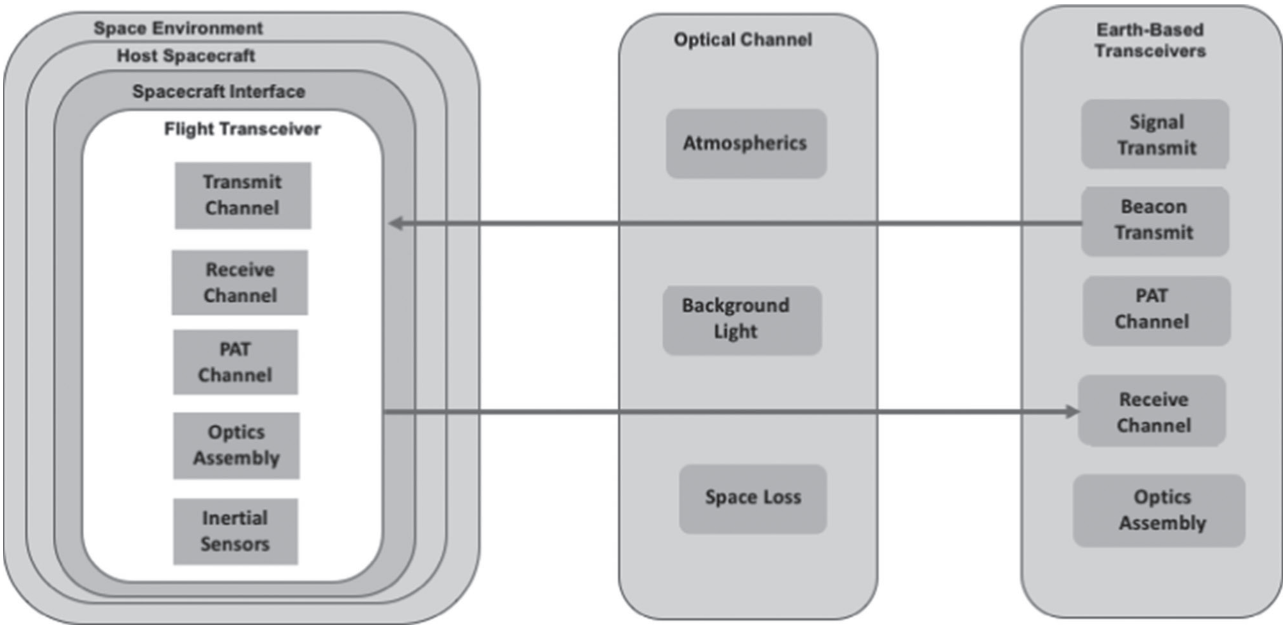


FIGURE 28.2 Free-space laser communications signal flow diagram from and to transceivers located at either the spacecraft or the ground station. Environment at either end of the link and the channel in-between can strongly influence the performance of the lasercom link.

For communication from spacecraft, typically the transceiver incorporates a dedicated PAT sub-system for stabilizing the line-of-sight. A reference beacon, such as a dedicated beacon laser or the transmit laser itself originating from the opposite communications transceiver, is typically required for satellite communications links since the distances are large and a cooperative target will help achieve the micro-radian-level beam-pointing requirement. Since the incoming beacon signal is faint, for a common transmit/receive aperture system effective isolation is required to detect the beacon signal void of the clutter from the scattered and any other unwanted signal originating from the much higher power transmit laser. A control loop with high bandwidth relative to the disturbance spectrum can also help in compensating for the platform jitter. The following steps are commonly implemented to precisely point beams generated by a lasercom transceiver situated on a platform that experiences:

Spatial Acquisition. To accomplish a true communication line-of-sight, the spatial-acquisition process's function is to compensate for large pointing errors at the onset of the link establishment. A typical link acquisition process starts with the transmitter (spiral or raster) scanning its low-divergence beacon or transmit laser beam over the uncertainty area until the receiver detects the incoming signal and consistently locks onto the signal. Degree of the knowledge of the satellites' position derives the uncertainty field-of-regard. For the most efficient process, it is important to minimize the acquisition time.

Tracking. The main function of the tracking assembly is to mitigate the host-platform's dynamic angular disturbances perturbing the lasercom transceiver's assembly. For lasercom with aircraft or spacecraft, the typical tracking accuracy is on the order of 1/10th the optical system's beamwidth.

Beam Pointing and Stabilization. At this stage the line-of-sight of the opposing lasercom transceivers is locked and tracking each other, while the two platform's relative cross-velocities and any vibration-induced residual disturbances are mitigated.

Sources of platform vibrations include reaction wheels, thruster firings, engaged motors, momentum dumping, solar panel actuation, target acquisition slewing and thermal gradients due to sun exposure that affects the opto-mechanical assembly. The magnitude of micro-vibrations could be in the range of 0.1 Hz to several 100 Hz with amplitude fluctuations larger than the transmit beam-width. Under these conditions, the sub-μrad pointing goal becomes highly challenging.

Jitter and bias error elements comprise the total boresight mis-pointing error, which is the result of misalignment between the transmit and receive axes of the optics assembly. High-frequency vibrational-mode frequency disturbances (jitter) are calculated via root-sum-square of its different constituents, while the low-bandwidth disturbance (bias) elements are added linearly [22]. The 3-sigma total pointing accuracy is then calculated as 3x jitter+bias. Contributing parameters to the overall beam-pointing error of a flight lasercom transceiver include residual (uncompensated) tracking errors, boresight calibration errors, position determination errors, a variety of sensor noises, boresight alignment errors and point ahead/behind calculation error (when applicable). Both passive and active vibration isolators may be applied to partially reject host-platform disturbances that are imparted onto the lasercom transceiver.

Depending on the magnitude of angular variations, a coarse pointing assembly (CPA) may be required in addition to the fine-pointing assembly (FPA). CPAs typically incorporate a 2-axis gimbal with stepper or brushless DC motors that are of high-enough precision and rated for space use. The FPA typically comprises one or more fine-steering mirror (FSM) to handle the high-frequency and acceleration portions of pointing mechanism and compensate for the residual vibrations and a focal plane (quadrant detector or a detector array) that measures the angular arrival variations based on the incoming beacon signal. Both voice-coil-actuated and piezo-based FSMs may be applied depending on the required angular range from the FSM [23]. Incorporating an inertial reference sensor that is tightly coupled with the flight transceiver's optics assembly can also help reduce the required centroiding update rates.

Figure 28.3 schematically illustrates a representative laser-com flight transceiver block diagram, identifying key PAT elements that work in coordination to precisely point the transmit laser beam to the intended receiver. The transceiver consists of a single telescope for a common transmit and receive path, an optical bench assembly (not shown) that encloses the transmit and receive beam-shaping and routing optics, the PAT mechanisms and the PAT sensors, a two-axis coarse-pointing gimbal assembly with a wide field-of-regard, and interface electronics co-located with the optical head.

28.2.2 Rugged Opto-Mechanical and Thermo-Mechanical Optics

The transmit and receive optics channels which may or may not share a common aperture (telescope), the aft optics channel (the optics bench), possibly a beacon transmit and receive channel, the acquisition and tracking channel, possibly the inertial sensors channel, and the mechanical structure that supports the optics, actuators, and the senor elements comprise the opto-mechanics assembly.

The optics assembly often has to incorporate a variety of field-of-views (FOVs) and a variety of wavelengths. Maintaining high wavefront quality over the entire FOV for the optics assembly (high Strehl ratio), adequate background and stray light rejection, and a structure with high fundamental frequency and high temporal and thermal stability are among the deriving requirements for the opto-mechanics assembly.

To meet the stringent (µrad level) beam pointing accuracy requirements, design and development of an optical-head assembly (OHA) with high temporal and thermal assembly in its opto-mechanical structure is of utmost importance. Moreover, the opto-mechanical system's structural design has to tolerate the spacecraft launch loads without suffering from misalignments. Driving OHA requirements may include large aperture diameter telescope (trades against the required laser output power) while maintaining low overall size and mass; capability to support multiple wavelengths (refractive vs. reflective design); adequate background and spurious light rejection; and high fundamental frequency for its structural stability. A variety of materials are available for the opto-mechanics structure, mirror substrates and optics support to satisfy flight terminal's mass constraints or thermal loading. Examples include SiC, Be and AlBeMet.

The use of fibre optics in the transmit and receive channels helps to minimize drift between these channels and allows for the use of an array of components advanced by the fibre

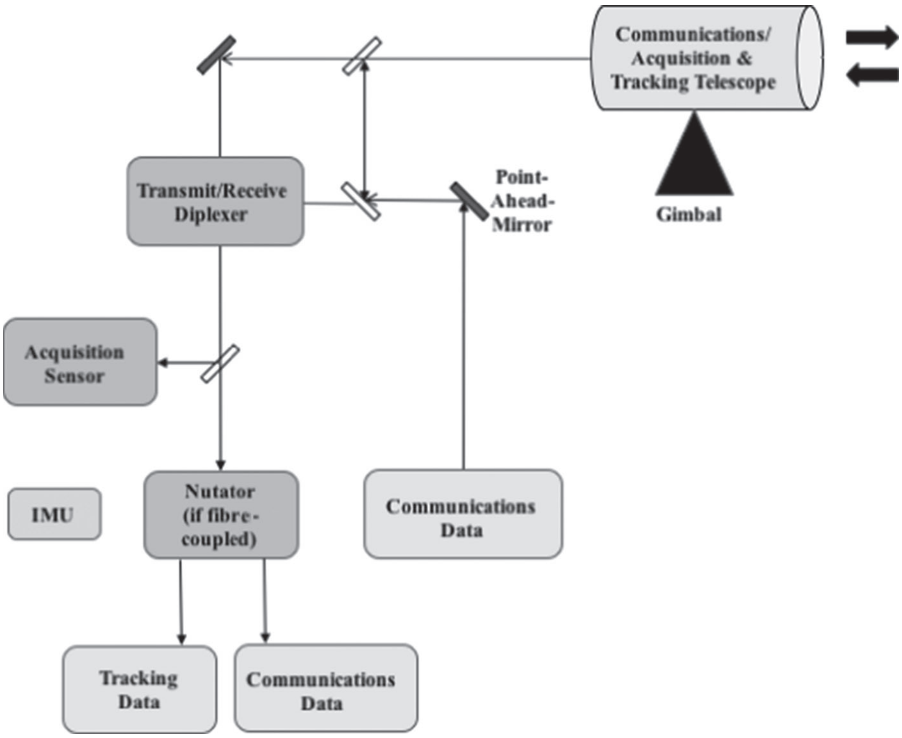

Communications/ Acquisition & Tracking Telescope
Gimbal
Transmit/Receive Diplexer
Point-Ahead-Mirror
Acquisition Sensor
IMU
Nutator (if fibre-coupled)
Communications Data
Tracking Data
Communications Data

FIGURE 28.3 A generic pointing/acquisition/tracking block diagram showing its key components. Here, the optical train incorporates a common path transmit/receive aperture, followed by transmit, receive and boresight channels. The receive channel might accommodate both data and beacon signals. The beacon signal originating from the opposite terminal is imaged on to a quadrant detector or a focal-plane-array detector for spatial acquisition and tracking and is used to aid with precision transmit-beam pointing. Laser transmitter's beam is reflected off of a two-axis fine-pointing-mirror prior to exiting the telescope and propagation through vacuum or air. For reference, to track the outgoing laser transmitter, a fraction of the outgoing beam might be reflected back onto the acquisition or tracking focal planes.

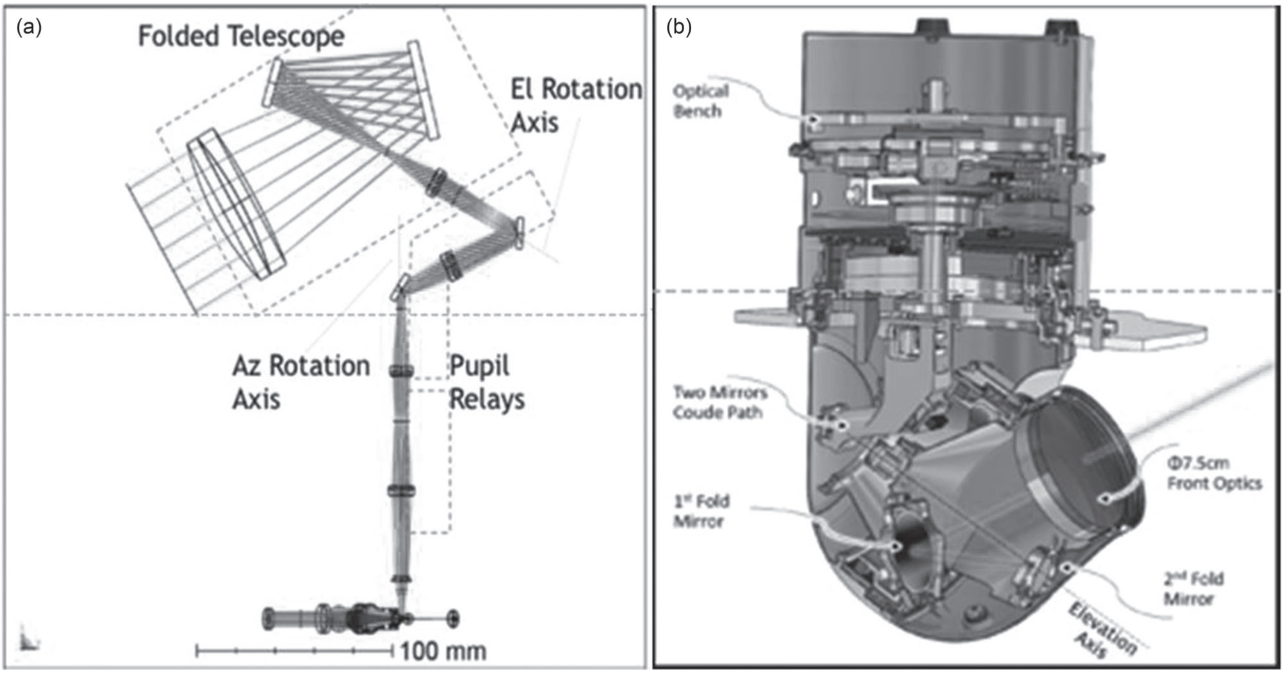


FIGURE 28.4 Optical layout of the gimbal optical module: (a) gimbal and (b) integrated into the mounting structure. (Credit: Facebook Inc.)

optics industry [24]. Launching a beam has traversed through the atmosphere from the satellite to ground requires an AO system incorporated with the ground station. However, since the wavefront of the uplink laser beam is adequately flat by the time it reaches the satellite, there is no issue in fibre-coupling at the satellite side.

Figure 28.4 illustrates an example of highly compact opto-mechanics design where the optical assembly is intimately combined with the coarse-pointing-assembly. A folded refractor 7.5 cm refractor telescope is followed by a Coudé path that relays the optical signal at the pupil plane onto a small folded-path optics bench housing the relay optics, the narrow and the wide-FOV optical tracking sensors, filters as well as a number of FSMs to couple the received beam with a single-mode fibre [25].

The optics bench uses a common-mode FSM to stabilize the incoming line-of-sight (Figure 28.5). A beamsplitter separates out the beacon signal reflected off the FSM from the communication signal since they are separated in wavelength by about 20 nm. Another FSM is used in conjunction with a focusing lens to steer the communications beam onto a single-mode fibre. Focusing lenses for the quadrant detector and fibre-coupling are mounted on miniature motorized translation stages to allow for fine position adjustments that may arise as a result of the extreme thermal environment [26].

28.2.3 Signal Detection

The signal detectors on both ends of the link convert the optical signal into electrical signals for further thresholding (to reject noise), amplification and processing (e.g. decoding, demodulation and error correction). Photon-starved communications regimes would require specific detectors (e.g. single-photon sensitive or photon-counting detectors). Receiver sensitivity ultimately limits the achievable link performance (low BERs) [27].

Figure 28.6 schematically illustrates one of many implementation options in the data transmit and receive chains. Commonly, transmit data along with coding and a specific modulation format are digitally applied to a low-power oscillator (laser) via direct or external modulation. A fibre power amplifier (e.g. Erbium-Doped fibre amplifier—EDFA) amplified the signal to the required level prior to propagation through the optics assembly. At the receiver side, the optical signal is converted to electrons followed by amplification and decoding/demodulation to extract the transmitted signal.

Both coherent and incoherent (intensity-modulated direct-detection) modes of operation are available to laser communications. Coherent detection schemes conventionally applied to RF and micro-valve communications have reached a high degree of maturity owing to the major advancements in the fibre optics communications industry. Compact and lightweight 400 Gbps (and soon 1.2 Tbps) transceivers are now commercially available. A number of satellite link demonstrations are now under development in 2019 using the coherent transceivers developed for the fibre optics industry. A challenge posed by coherent links that traverse the atmosphere is that an AO system has to be incorporated with the ground station for phase-front correction so that the incoming signal may be coupled back into a single-mode fibre. Coherent detection allows the use of binary-phase-shift-keying modulation and requires transmission of laser signals with both single spatial and longitudinal mode.

Pre-amplified coherent receivers can provide performance characteristics that approach the theoretical limits [28,29]. For applications where the photodetector remains in linear mode, coherent receivers are capable of very effectively rejecting the background light and that is of major advantage where at

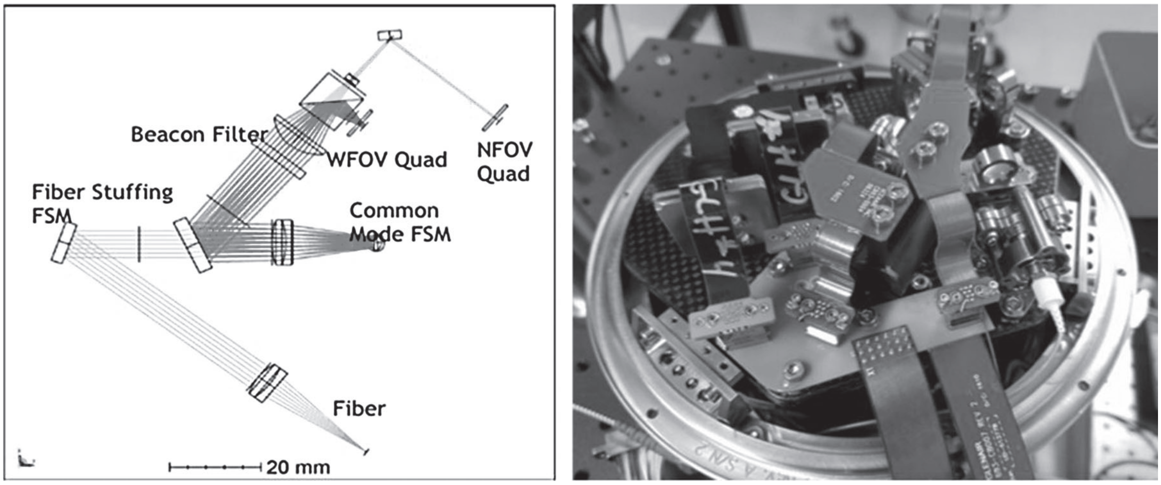

Beacon Filter
WFOV Quad
NFOV Quad
Fiber Stuffing FSM
Common Mode FSM
Fiber
20 mm

FIGURE 28.5 The small optics bench for the aerial lasercom link where large number of optics, sensors and actuators are packaged in a 10 cm diameter area. (Credit: Facebook Inc.)

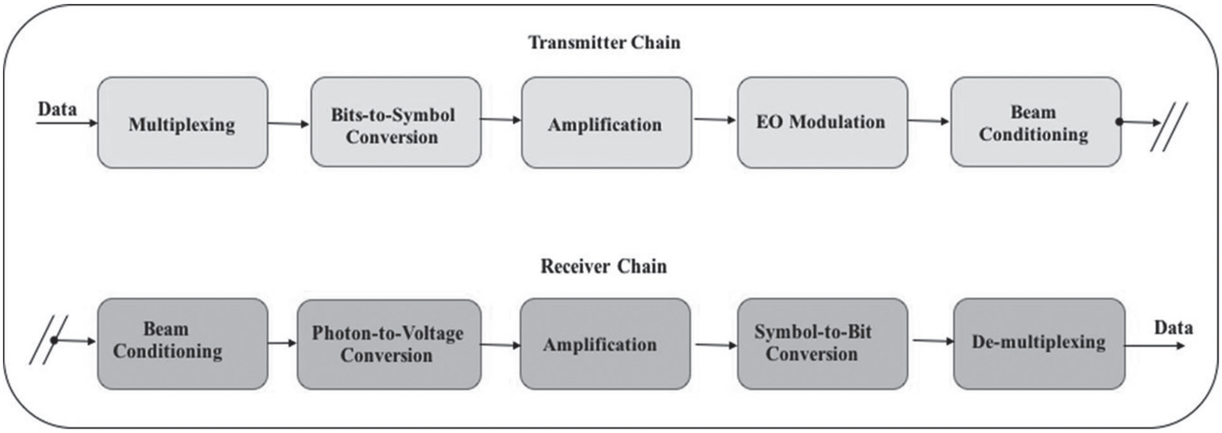

Transmitter Chain
Data
Multiplexing
Bits-to-Symbol Conversion
Amplification
EO Modulation
Beam Conditioning
Receiver Chain
Beam Conditioning
Photon-to-Voltage Conversion
Amplification
Symbol-to-Bit Conversion
De-multiplexing
Data

FIGURE 28.6 A high-level block-diagram representation of one of several approaches to free-space laser communications signal transmit and receive.

occasions sunlight directly accompanies the communications link. The maximum achievable coherent-detection photon efficiencies are1.44 bits/photon for heterodyne detection and 2.89 bits/photon for homodyne detection [30,31]. Leveraging the WDM and the highly sensitive receivers developed for the fibre optics industry can make possible cost-effective satellite-based laser communications.

Both PIN- and APD-type photodetectors are available for use in lasercom with Si, Ge, InGaAs and HgCdTe detectors [32]. Each of these materials has different responsivity and noise characteristics making them suitable for a specific communications application. Ge is used less frequently, while HgCdTe is finding more application. Super-conducting nanowire arrays and Geiger-mode detectors are options in power-starved regimes [33,34].

28.2.4 Modulation and Coding

The technology of digitally superimposing the (transmit) data onto the laser beam is referred to as modulation. To prevent or to correct for signal transmission errors, specific error-correcting codes that match a particular link may be simultaneously implemented along with modulation. Today's highly efficient modulation and coding schemes have approached to within a fraction of 1-dB of Shannon limit [35]. Yet, more improvements are still anticipated. The broad frequency bandwidth of the optical band enables the use of advanced bandwidth-sensitive modulation and coding schemes to deliver higher data rates and to simplify implementation.

Given a receiver's architecture, Shannon's classical treatment of information theory sets capacity limits on the highest reliability of data communications, while Holevo's quantum-mechanical treatment of information theory optimizes link capacity over the best possible receiver [36]. Holevo's theorem provides predictions of an optical receiver optimized via quantum mechanics to achieve quantum-limited detection. The latter has not completely materialized experimentally yet.

Photon and power-efficient implementation is often a key design-driver for modulation and waveform generation hardware. The uncoded OOK (on-off-keying) modulation format is a low-complexity and high-performance format for lasercom links that are not severely sensitivity- and capacity-constrained

[37]. For moderately constrained lasercom links, where sensitivity requirements are moderately strict, the PSK modulation scheme provides strong performance for lasercom links [38]. Binary differential phase-shift keying has also been used to achieve high-sensitivity and high-data-rate lasercom with optically pre-amplified receivers [39,40]. Low-density parity check codes and the efficient parallel or serially-concatenated PPM codes are among some the most efficient coding schemes [41].

FEC codes can correct receiver's detection errors and provide redundancy at the transmitter [42]. Sensitivity gains (improved link margin) of as much as 10 dB with <10% overhead but with additional implementation complexity may be achieved at data rates of tens of Gbps [43]. Commercially available FECs based on Reed-Solomon codes are available at rates up to 40 Gbps and with ~7% overhead.

28.2.5 Flight Laser Transmitter

The laser transmitter of choice for satellite laser communications is often of the master-oscillator-power-amplifier (MOPA) type since it can provide moderate-to-high average power levels with high output beam quality, low jitter, high fidelity, high MER and at data rates that can span to multi-Tbps. Such optical amplifiers may also be applied for pre-amplification in front of the receiver [44]. A highly simplified MPA configuration is shown in Figure 28.7. The optical amplifier boosts the power of the input laser with gains of tens of dB while preserving the modulation quality of the input (master oscillator) laser.

A continuous-wave distributed-feedback (DFB) laser typically constitutes the master-oscillator. Yb-doped and Er-doped fibre amplifiers (YDFA and EDFA) with wavelengths on at approximately 1000 nm and 1550 nm are most commonly used, although a variety of other fibre-doped amplifiers covering the 1000–1700 nm wavelength range are available. Wall-plug efficiencies are modest, in the order of 15% and 20% at the beginning of life for EDFA and YDFA, respectively [45].

Dispersion effects on the fibre, often an issue on terrestrial communications applications, is not an issue for laser communications, but non-linear effects (at higher powers) and amplified spontaneous emission accompanying the transmit signal need to be avoided. Oscillator and amplifier components qualified for terrestrial fibre communications often meet majority of the flight qualification requirements (e.g. vibration, shock and lifetime), but not operation under the vacuum and cosmic radiation environment. Many such lasers have been qualified for space and have performed well in the space environment [46].

28.2.6 The Atmospheric Channel

The atmosphere has a number of adverse effects on the propagating laser beam, including:

- **Attenuation** through absorption via its constituent elements, and scattering off particles, e.g. fog/cloud, rain and snow
- **Scintillation/turbulence** via refractive index fluctuation effects, resulting in beam deflection (tilt and jitter) and phase degradation (wavefront aberrations).
- **Radiance/Background light** resulting in increased noise at the optical sensors.

Atmospheric zones with difference indices of refraction are created when sun-heated earth warms up the atmosphere in its vicinity. This creates prism-like regions that can substantially deviate a propagating laser beam and resulting in phase and irradiance fluctuations. The resulting turbulent atmosphere adversely affects the traversing laser beam by inducing angle-of-arrival fluctuations, spatial and temporal intensity fluctuations, beam steering, beam wander, speckle, image dancing at the focal plane of the receive optics, destructive interference (scintillation), beam focusing and beam defocusing (spreading)—either individually or simultaneously. The ensuing signal fading effects may occur on the order of hundreds of Hz to several kHz.

The Kolmogorov model provides a quantitative measure of the turbulent atmosphere via an *index of refraction structure parameter* called C_n^2 [47], via the following relationship:

$$C_n^2 r^{2/3} = \left[n(r1) - n(r2)\right]^2$$

where r is the distance over which the turbulence is measured and n is the index of refraction. Weak turbulence is on the order of $10^{-17}\mathrm{m}^{-2/3}$ and strong turbulence is on the order of $10^{-13}\mathrm{m}^{-2/3}$. A highly useful quantitative measure, called the Fried's parameter or the atmospheric coherence length parameter (r_0), is related to C_n^2 via [48]:

$$r_0 = \left(0.423 k^2 C_n^2 L\right)^{-3/5}$$

where $k = 2\pi/\lambda$ and L is the "slant" range to aerial or space-borne platforms. As examples, an $r_0 \approx 2$ represents a severely turbulent atmosphere (happening during daytime) and an $r_0 \approx 20$ signifies a low turbulent period (happens typically during night time). Strong turbulence is often characterized by

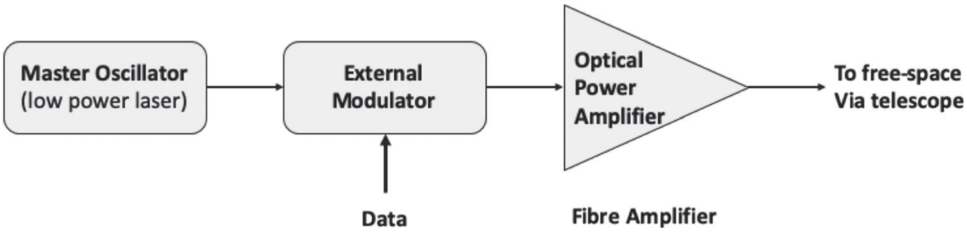

Master Oscillator (low power laser)
External Modulator
Data
Optical Power Amplifier
Fibre Amplifier
To free-space Via telescope

FIGURE 28.7 MOPA transmitter configuration, consisting of a modulated semiconductor laser (e.g. a DFB) and a rare-element-doped fibre (optical) power amplifier.

exponential statistics, while weak scintillation is often characterized by lognormal statistics. The Hufnagel-Valley theorem models atmospheric fluctuations, based on C_n^2, for different altitudes [49].

28.2.6.1 Mitigating Effects of Turbulence/Scintillation

28.2.6.1.1 Multi-Beaming for Uplink

It has been demonstrated analytically and empirically that the transmission of multiple laser beams that are coherently independent from each other and are separated from one to the other by as much as r_0 can reduce the received signal fluctuations by as much as $1/N$ for N up to four beams and by $1/N^{1/2}$ for $N>4$ beams [50]. Diminishing returns are seen with 16 or more beams. This phenomenon depends on different beams traversing different portions of the atmosphere, and in average experiencing less turbulence effects than a single beam propagating the atmosphere.

28.2.6.1.2 Aperture Averaging for Downlink

A receive ground optical aperture of diameter D captures $(D/r0)^2$ independent turbulence cells. "Aperture averaging" denotes an average intensity of different receive beams traveling through independent atmospheric cells. For the direct-detection technique, magnitude of the aperture averaging effects is directly proportional to the diameter of the ground telescope and becomes more pronounced for apertures beyond 30 cm in diameter [51]. For coherent detection, the beam wavefront maintains coherence over a distance r0, which limits the effective ground aperture diameter to r0 and without the use of AO technology a larger diameter telescope would not increase the level of signal coupled into a single mode.

28.2.6.1.3 AO Technique for Downlink and Uplink

The AO technique is used to undo the effects of atmosphere in the laser beam. This technique is now well-established for downlink beams, but in its infancy for uplink beams. The AO system incorporates a wavefront sensor that characterizes the quality of the incoming beam and a deformable mirror with pixelated and individually addressable mirrors that can be used as the wavefront sensor to correct for the incoming beam's aberrations [52]. The AO system in effect mitigates signal fades and therefore increases the link's SNR, thereby enhancing the link margin by as much as 6 dB. A highly useful benefit of AO to lasercom is that this technique makes efficient coupling of an atmospheric-propagated beam to a single-mode fibre (essential for coherent communications).

The uplink beam is more severely affected by scintillation than the downlink beam. Due to the point-ahead effect, the downlink/uplink channel is not reciprocal and downlink AO is ineffective for uplink. Approaches that could be effective include laser guide star, Rayleigh backscatter laser and an orbiting laser beacon. This technology is at the experimental stage at this point [53].

28.2.6.1.4 Coding of Downlink and Uplink Transmitters

Akin to the multi-beaming spatial-diversity approach described earlier, block codes are demonstrating potential in mitigating the turbulence effects for both cases of downlink and uplink beams [54]. Implementation of inter-leaver codes that distribute atmospheric fades over many code words of an error-correction code may mitigate the losses. Just like the multi-beaming approach that each beam experiences different fades; in the case of coding each code word encounters a number of uncorrelated fades. The resulting performance average is a major reduction in signal fades.

28.3 Ground Station

28.3.1 Ground Receive Telescope

The ground telescope subsystem tracks the air-borne or space-borne platform, collects the downlinked photons and processes the received signal following the rejection of background noise. Unlike astronomical telescopes that predominantly operate at night time, a lasercom ground station has to operate at all times and that creates a certain number of design challenges, such as extreme thermal stability and operations at small sun angles.

The turbulence-induced effects via the atmosphere just above the telescope degrade the receive optical assembly's effective optical quality. These effects may include random fluctuations of received laser power, angle-of-arrival fluctuation due to wavefront tip/tilt effects, wavefront distortions (blurring). Large diameter ground telescopes simplify the link by adding margin to the link. However, unlike RF/MW ground stations with multi-metre antenna diameters that are plentiful and relatively inexpensive, the number of lasercom ground stations is limited and mostly is in the aperture diameter range of 20 and 150 cm. Atmospheric monitoring instruments are often included with the ground station to correlate outage events with specific atmospheric-induced effects that may have occurred. Figure 28.8 shows pictures of the NASA-JPL's 1-m diameter Optical Communications Telescope Laboratory (OCTL), and Facebook's in California, and Facebook's 0.2-m diameter AO-equipped ground telescope, both located in the Los Angeles (CA) area.

A number of commercial vendors are now working on developing a global network of laser communications ground stations [55]. Availability of such ground networks would alleviate one of the major impediments to making possible a more prevalent use of laser satellite communications.

28.3.2 Ground Uplink Transmitter

The uplink communications and beacon laser (if any) could either be transmitted through the downlink receive telescope or via a dedicated optics assembly. Shared aperture would require adequate transmit/receive isolation. This may be accomplished via wavelength diversity or polarization. Compliance with regulatory bodies and eye-safety standards are often required to safely propagate high-power laser beam through the atmosphere up to airplanes or satellites. This may require addition of hazard detection equipment such as lidar and radar for airplane detection, and laser safety interlocks are automatically activated.

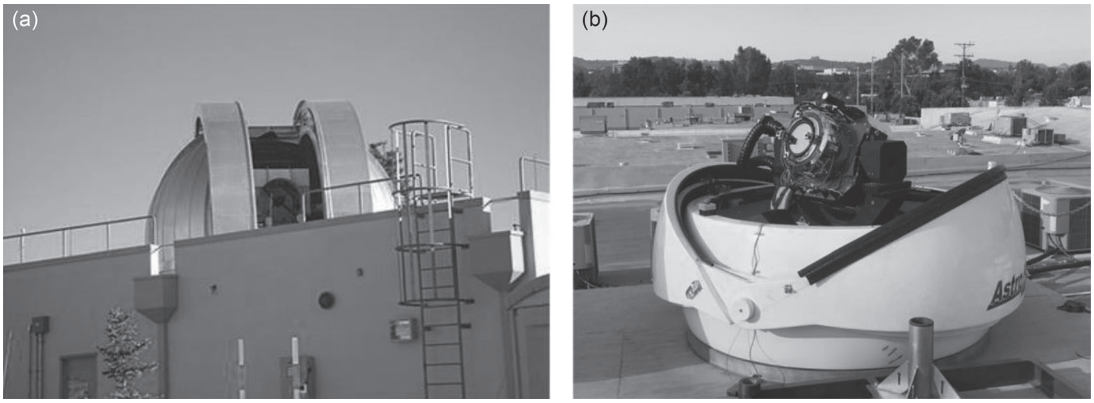


FIGURE 28.8 (a) NASA-JPL's 1-m diameter lasercom-dedicated telescope (called OCTL) capable of tracking LEO and GEO spacecraft and fast-moving aircraft. This telescope is capable of operating at daytime to within 10° of the Sun, and with an add-on filter, to less than 3° of the sun. (b) A 20-cm aperture diameter telescope equipped with AO is capable of tracking fast-moving aircraft and particularly suitable for coherent communications links.

28.3.3 Ground PAT

The PAT subsystem's role in the ground station is to lock on to the incoming downlink or beacon signal and to track the aerial or space platform within a pre-set time window. The telescope mount calibration has to be conducted in advance so that the telescope can blind-point to the coordinates or the ephemeris of the satellite. The acquisition/tracking assembly typically incorporates a wide-FOV and a narrow filed-of-view senor for acquisition and tracing, respectively, and at least one FSM to direct the downlink signal to within the FOV of the tracking sensor (quadrant detector or a larger detector array), and for atmospheric-induced tip-tilt correction [56]. The point-ahead function is typically accomplished via a separate FSM.

28.4 Applications

Satellite communications links not requiring real-time data-delivery, such as down-linking large volumes of science or sensor data are well served by high-data rate through-atmosphere lasercom links that can suffer from atmospheric outages. Similarly, defense-related links that require high level of security and immunity to jamming.

High link availability laser communications links include cross-links (e.g. LEO-LEO and LEO-GEO) and high-rate feeder links to GEO satellites since in the latter case station diversity may be applied to circumvent atmospheric outages. Station diversity would require a number of stations that are located in independent weather cells (generally over 400 km apart) and in sufficient number to be within the FOV of the satellite optics assembly. LEO-LEO cross-links are of significant interest at the time of this writing since a number of LEO satellite constellation providers are relying on this technology to minimize the required number of ground gateway stations to relay data from one point on earth to another [57]. Flight-qualified lasercom transceiver delivery challenges are meeting the tight constraints of cost, mass, power and quantity delivery (usually four transceivers per satellite) in short time period.

Rapid advances by the fibre optics communications community and the photonic-integrated circuit technology is accelerating the pace of advances for the free-space lasercom technology [58,59]. More efficient receivers, amplifiers, modulation and error correction, and simplified PAT architectures along with improved coarse-pointing capability are expected to relax the flight lasercom transceivers that constitute the bulk of a link's cost. For example, there is little to gain in most scenarios with further theoretical gains in coding but huge gains feasible in its implementation.

Other applications include the following:

- satellite networking allowed by the large number of LEO satellite constellations that are now under development [60];
- point-to-multi-point communications, a technology that it is still maturing even after decades of development [61];
- superior navigation ranging, a by-product of flight communications transceiver, made possible by the short wavelength of laser beams (relative to that of RF/MW beams) [62];
- multi-functional transceivers, by combining multiple active and passive optical functions (e.g. lidar and imaging) with a lasercom terminal [63];
- retro-modulator links for niche applications where unlink typical communications systems, the transmitter and the receiver are co-located. Data rates on the order of 10 Mbps and ranges of several kilometres have been demonstrated [64];
- hybrid-RF/optical links that are suitable for applications that rely on high link availability is a requirement. The RF/MW and lasercom technologies could be used as a back-up to each other to extend the availability of the link and at the same time to maximize the link's capacity. A number of such links have been demonstrated in the past [65];
- deep-space (planetary range) communications system development is now underway at JPL/NASA

from an asteroid [66]. The Lunar range is the farthest that lasercom has communicated so far [67,68]. For comparison's sake, due to the $1/R^2$ effect, communicating with a spacecraft at 30 AU (astronomical unit) is 10 billion times more difficult than communicating with a satellite at GEO (~40,000 km away). Received signal level from deep space is on the order of tens of fW/m^2 vs. a few mW/m^2 for near-earth links. Also, deep-space beam-pointing requirements are on the order of sub-micro-radian vs. several micro-radians for earth-orbiting links with round-trip light times that could be on the order of hours vs. seconds for near-earth links.

The RF satellite community is making major strides in developing the technology of high throughput satellites (HTS) to the point that operation Tbps class Ka-band satellite is expected by 2022 [69,70]. The total capacity is formed by hundreds to thousands of downlink or uplink beams each at 1–2 Gbps data rate. With the ever-increasing and near-exponential demand for data capacity, as experienced by the terrestrial fibre optics communications, the lasercom technology can augment the RF HTS satellites, for example, via Tbps uplink to the satellites or hundreds of Gbps cross-link between satellites in LEO, MEO or GEO. Future technology advancements are anticipated through simplified and lower-cost sub-systems, in many cases taking advantage of development in related electro-optic technologies. Examples of flight-qualifiable sub-systems include higher wall-plug optical amplifiers, low-cost optical assemblies that are amenable to rapid integration and test, and higher detection efficiency and lower noise photodetectors. On the ground side, there is a need for much-reduced cost metre-class ground telescopes that meet laser communication's demanding requirements.

REFERENCES

1. Cisco 2016, "Cisco global cloud index: Forecast and Methodology, 2015–2020" White paper.
2. D.K. Sachdev, Published Online, 2017. doi:10.2514/4.104053.
3. C. Miller, "How and why commercial high-capacity satellites offer superior performance and survivability in the future space threat continuum," *32nd Space Symp.*, Technical Track, Colorado Springs, Colorado, United States of America. Presented on April 11–12, 2016.
4. M. Zhang, C. Wu, Y. Jiao, S. Zhai and Y. Zhao, "Multi-beam reflector antenna architecture for beam-hopping and size reduction system," *2017 Sixth Asia-Pacific Conf. on Antennas and Propagation (APCAP)*, Xi'an, 2017, pp. 1–3.
5. H. Hemmati, *Near-Earth laser Communications*, CRC press, Boca Raton, FL (2009).
6. H. Hemmati, *Deep Space optical Communications*, Wiley (2006).
7. I. del Portillo, B. Cameron, and E. F. Crawley, "A technical comparison of three low earth orbit satellite constellation systems to provide global broadband," *Acta Astonautica*, V. 159, 123–135 (2019).
8. P. M. Pimentel, N. Hoepcke, C. Seiter, K. Saucke, F. Heine, M. Motzigemba, and S. Martin, "For years operations of inter-satellite (ISLs) and space-grpund (SGLs) optical links," *2018 SpaceOps Conf.*
9. H. Hemmati and D. Caplan, *"Optical Satellite Communications,"* Optical Fiber telecommunications Volume V1B, Chapter 4, 6th edition by A. E. Wilner, T. Li, and I. Kaminow (2014).
10. Y. Arimoto, M. Toyoshima, M. Toyoda, T. Takahashi, M. Shikatani, and K. Araki, "Preliminary result on laser communication experiment using ETS-VI", *Proc. SPIE*, vol. 2381, San Jose, California, pp. 151–158 (1995).
11. K. E. Wilson et al., "Overview of the ground-to-orbit lasercom demonstration," *Free-Space Laser Communication Technologies, IX, Proc. SPIE*, V. 2990 (1997).
12. K. Nakagawa and A. Yamamoto, "Engineering model of LUCE (Laser Utilizing Communications Experiment)," *Proc. SPIE*, V. 2699 (1996).
13. R. Fields, D. Kozlowski, H. Yura, R. Wong, J. Wicker, C. Lunde, M. Gregory, B. Wandernoth, and F. Heine, "5.625 Gbps bidirectional laser communications measurements between the NFIRE satellite and an Optical Ground Station," *Proc. of Int. Conf. on Space Optical Systems and Applications*, pp. 44–53 (2011).
14. H. Kaempfer et al., "The alphasat GEO laser communication terminal flight acceptance tests," *Proc. of ICSOS* (2012).
15. D. M. Boroson, B. S. Robinson, D. V. Murphy, D. A. Burianek, F. Khatri, J. M. Kovalik, Z. Sodnik, and D. M. Cornwell, "Overview and results of the lunar laser communication demonstration," *Proc. of SPIE*, V. 8971, Free-Space Laser Communication and Atmospheric Propagation XXVI (2014).
16. chrome-extension://oemmndcbldboiebfnladdacbdfmadadm/https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-451-principles-of-digital-communication-ii-spring-2005/video-lectures/chap4.pdf.
17. D. M. Boroson, "A survey of technology-driven capacity limits for free-space laser communications," *Proc. SPIE*, V. 6709, Free-Space Laser Communications VII (2007).
18. J. Kauffmann, "Performance limits of high-rate space-to-ground optical communications through the turbulent atmospheric channel," *Proc. SPIE*, V. 2381, pp. 171–182 (1995).
19. L. C. Andrews, R. L. Phillips, *Laser Beam Propagation Through Random Media*, 2nd Edition, SPIE Pub. (2005).
20. R. M. Gagliardi and S. Karo, *Optical Communications*, 2nd Edition, Wiley (1995).
21. L. B. Stotts, P. Kolodzy, A. Pike, B. Graves, D. Dougherty, and J. Douglass, "Free-space optical communications link budget estimation," *Applied Optics*, V. 49, Issue 28, 5333–5343 (2010).
22. R. J. Watkins, B. N. Agrawal, Y. S. Shin, and H.-J. Chen, "Jitter control of space and airborne laser beams," *22nd AIAA International Communications Satellite Systems Conf. & Exhibit 20049*, Monterey, California (2004).
23. Y. Kaymak, et al., "A survey of acquisition, tracking, and pointing mechanisms for mobile free-space optical communications," *IEEE Communications Surveys & Tutorials*, V. 20, Issue 2, 1104–1123 (2018).
24. E. A. Swanson and R.S. Bondurant, "Using fiber optics to simplify free-space lasercom systems," *Proc. SPIE*, V. 1218 (1990).
25. C. Chen et al., "Demonstration of bi-directional coherent air-to-ground optical link," *Proc. of SPIE Photonics West*, V. 10524 (2018).

26. C. Chen et al., "High speed optical links for UAV applications," *Proc. of SPIE Photonics West*, V. 1009615 (2017).
27. D. O. Caplan, "Laser communication transmitter and receiver design," *Journal of Fiber Communications Reports*, V. 4, 225–362 (2007).
28. M. S. Erkilinc, D. Lavery, K. Shi, B. Thomsen, R. Killey, S. Savory, and P. Bayvel, "Comparison of low-complexity coherent receivers for UDWDM-PONs (l-to-the-user)," *Journal of Lightwave Technologies*, V. 36, 3453–3464 (2018).
29. D. J. Geisler, C. Schieler, T. Yarnall, M. Stevens, B. Robinson, and S. Hamilton, "Demonstration of a variable data-rate free-space optical communication architecture using efficient coherent techniques," *Optical Engineering*, V. 111605, 111605 (2016).
30. B. Erkmen, B. Moision, and S. Dolinar, "On approaching the ultimate limits to resource efficiency in photonics communication," *Proc. of SPIE Photonics West*, V. 8246 (2012).
31. S. Dolinar, B. Erkmen, B. Moision, K. Birnbaum, and D. Divsalar, "The ultimate limits of optical communication efficiency with photon-counting receivers," *Proc. IEEE Int. Symp. on Information Theory*, 1–6 July (2012).
32. J. Rothman, G. Lasfargues, and J. Abergel, "HgCdTe APDs for free space optical communications," *Proc. SPIE*, V. 9647 (2015).
33. W. A. Stern and W. H. Farr, "Fabrication and characterization of superconducting NbN nanowire single photon detectors," *IEEE Transactions on Applied Superconductivity*, V. 117, 306–309 (2007).
34. C. Langrock, E. Diamanti, R. Roussev, Y. Yamamoto, M. M. Fejer, and H. Takesue, " Highly efficient single-photon detection at communication wavelengths by use of upconversion in reverse-proton-exchanged periodically-poled $LiNbO_3$ waveguides," *Optics Letters*, V. 30, 1725–1727 (2005).
35. B. Moision and J. Hamkins, "Coded modulation for the deep space optical channel: serially concatenated PPM," *JPL's IPN Progress Report*, V. 42, Issue 161, 1–26 (2005).
36. D. M. Boroson, "Toward optimum efficiency in a quantum receiver for coded PPM," *SPIE Proc. of Int. Conf. on Space Optics (ICSO)*, V. 10562 (2016).
37. D. M. Boroson, "A survey of technology-driven capacity limits for free-space laser communications," *Proc. SPIE*, V.6709, 670918 (2007).
38. J. Li, J. Q. Liu, and D. P. Taylor, "Optical communication using subcarrier PSK intensity modulation through atmospheric turbulence channels," *IEEE Transactions on Communications*, V. 55, 1598–1606 (2007).
39. A. O. Aladeloba, A. J. Phillips, M.S. Woolfson, "Improved bit error rate evaluation for optically pre-amplified free-space optical communication systems in turbulent atmosphere," *IET Optoelectronics* V. 6, 26–33 (2012).
40. D. J. Geisler, T. M. Yarnall, M. L. Stevens, C. Schieler, B. S. Robinson, and S. Hamilton, "Multi-aperture digital coherent combining for free-space optical communication receivers," *Optics Express*, V. 24, 12661–12671 (2016).
41. J. Hamkins and B. Moision "Selection of modulation and codes for deep-space optical communications," *Proc. SPIE*, V. 5538 (2004).
42. S. Yamamoto, H. Takahira, and M. Tanaka "5 Gbit/s optical transmission terminal equipment using forward error correcting code and optical amplifier," electronics let., V. 30, 254–255 (1994).
43. N. Chand, J. J. Loriz, J. Hunton, and B. Eleso, "Performance comparison of NRZ and RZ modulations with and without forward error corrections for free-space optical communication," *Proc. SPIE Optics and Photonics*, V. 5892 (2005).
44. M. J. F. Digonnet, *Rare-Earth-Doped Fiber Lasers and Amplifiers*, CRC Press, Boca Raton, FL (2001).
45. I. Lavrinovica and J. Porins, "An improvement of EDFA efficiency by using co-doped optical fibers," *2016 IEEE International Black Sea Conf. on Communication and Networking*, 6–9 June (2016).
46. A. Wu et al., "Overview of space qualified solid state lasers development at NASA Goddard Space Flight Center," *Proc. SPIE Solid State Lasers* V. 7193 (2009).
47. D. Gulich, G. Funes, D. Perez and L. Zunino, "Estimation of C_n^2 based on scintillation of fixed targets imaged through atmospheric turbulence," *Optics Letters*, V. 40, 5642–5645 (2015).
48. S. Basu, "A simple approach for estimating the refractive index structure (C_n^2) profile in the atmosphere," *Applied Optics*, V. 40, 4130–4133 (2015).
49. J. L. Bufton, "Comparison of vertical profile turbulence structure with stellar observations," *Applied Optics*, V. 12, 1785–2793 (1973).
50. I. I. Kim, H. Hakakha, P. Adhikari, E. Korevar, and A. Majumdar, "Scintillation reduction using multiple transmitters," *Proc. SPIE*, V. 2990, 102–112 (1997).
51. J. H. Churnside, "Aperture averaging of optical scintillations in the turbulent atmosphere," *Applied Optics*, V. 30, 1982–1994 (1991).
52. R. Tyson, "Bit-error rate for free-space adaptive optics laser communications," *Journal of the Optical Society of America*, V. 19, 753–758 (2002).
53. N. Martinez, L. F. Rodriguez-Ramos and Z. Sodnik, "Toward the uplink correction: application of adaptive optics techniques on free-space optical communications through the atmosphere," *Optical Engineering*, V. 57, 076106, (2018).
54. R. Pernice, A. Ando, A. Parisi, A. C. Cino, and A. C. Busacca, "Moderate-to-strong turbulence generation in a laboratory indoor free space optics link and error mitigation via RaptorQ codes," *18th IEEE Int. Conf. on Transparent Optical Networks (ICTON)*, 10–14 July (2016).
55. E. Herz, C. Dahn, E. Carney, and J. Campagna, "BridgeSat laser communication scheduling: a case study," *2018 SpaceOps Conf.* (2018).
56. T. S. Rose, S. W. Janson, S. D. La Lumondiere, N. Werner, D. H. Hinkley, D. W. Rowen, R. A. Fields, and R. P. Welles, "LEO to ground optical communications from a small satellite platform," *Proc. SPIE*, V. 9354, Free-Space Laser Communication and Atmospheric Propagation XXVII (2015).
57. J. Alvarez and B. Wells, "Constellations, clusters, and communication technology: expanding small satellite access to space," *2016 IEEE Aerospace Conf.*, 5–12 March (2016).
58. "Widely tunable laser transmitter for free space optical communications," *2018 IEEE International Semiconductor Laser Conf. (ISLC)*, 16–19 September (2018).
59. J. Klamkin, H. Zhao, B. Song, Y. Liu, B. Isaac, S. Pinna, F. Sang, and L. Coldren, "Indium phosphide photonic integrated circuits; technology and applications," *2018 IEEE BiCMOS and Compound Semiconductor Integrated Circuits and tch. Symp.*, 15–17 October (2018).

60. V. W. S. Chan, "Free Space Optical Communications," *Journal of Lightwave Technologies*, V. 24, 4762 (2006).
61. S. W. Sparrold, E. L. Upton, and A. Okorogu, "Free space point-to-multipoint optical communication system and apparatus," U.S. patent US6912360 (28 June 2005).
62. Y. Chen, K. Birnbaum, and H. Hemmati, "Active laser ranging over planetary distances with millimeter accuracy," *Applied Physics Letters*, V. 102, 241107 (2013).
63. H. Hemmati, "Combined laser communications and laser ranging for Moon and Mars," *Proc. SPIE*, V. 7199 (2009).
64. W. S. Rabinovich, R. Mahon, P. Goetz., E. Waluschka, D. S. Katzer, S. Binari and G. C. Gilbreath, "A cat's eye multiple quantum well modulating retro-reflector," *IEEE Photonics Technology Letters*, V. 15, 461–463 (2003).
65. L. B. Stotts, et al., "Hybrid optical RF airborne communications," *IEEE Proceedings*, V. 97 (2009).
66. A. Biswas, "Status of NASA's deep space optical communications technology demonstration," *2017 IEEE Int. Conf. on Space Optical Systems and Applications (ICSOS)*, 14–16 November (2017).
67. D. V. Murphy, J. E. Kansky, M. E. Grein, R. T. Schulein, M. M. Willis, and R. E. Lafon, "LLCD operations using the lunar lasercom ground terminal," *Proc. of SPIE*, V. 8971, Free-Space Laser Communication and Atmospheric Propagation XXVI (2014).
68. B. S. Robinson, A. J. Kerman, E. A. Dauler, R. J. Barron, D. O. Caplan, M. L. Stevens, J. J. Carney, S. A. Hamilton, J. K. W. Yang, and K. K. Berggren, "781-Mbit/s photon-counting optical communications using a superconducting nanowire detector," *Optics Letters*, V. 31, Issue. 4, 444–446, 2006.
69. A. Karasuwa, J. Eastman, and I. Otung, "Design considerations for high-throughput satellite communication system," *Proc. 21st Ka and Broadband Satellite Conf.* (2015)
70. B. Roy, S. Poulenard, S. Dimitrov, R. Barrios, D. Giggenbach, A. Le kemec, and M. Sotom, *2015 Int. Conf. on Space Optical Systems and Applications (ICSOS)*, 26–28 Oct (2015).

29

Passive Silicon Photonic Integrated Components and Circuits for Optical Communications

Daoxin Dai, Yiwei Xie, and Yaocheng Shi

CONTENTS

29.1 Introduction 397
29.2 Silicon Nanophotonic Waveguides 398
29.3 Silicon Photonic Devices for On-Chip Polarization-Handling 400
  29.3.1 Silicon PBSs 400
  29.3.2 Silicon PRs 401
29.4 Multi-mode Silicon Photonics 402
  29.4.1 On-Chip Mode (De)Multiplexers 403
  29.4.2 Sharp Multi-mode Waveguide Bends 405
29.5 Wavelength-Selective Silicon Photonic Devices 406
  29.5.1 AWG-Based Optical Filters 406
  29.5.2 MRR-Based Optical Filters 407
    29.5.2.1 MRR-Based Optical Filters with Large FSRs 408
    29.5.2.2 MRR-Based Optical Filters with Box-Like Responses 408
    29.5.2.3 Polarization-Selective MRR Optical Filters 409
  29.5.3 Grating-Based Optical Filters 410
29.6 Reconfigurable Silicon Photonic Devices and Circuits 410
  29.6.1 Optical Switches 410
  29.6.2 ROADMs 412
29.7 Conclusion and Future Trends 414
References 415

29.1 Introduction

Integrated photonics have experienced a rapid advance in the past few decades, and various materials/platforms for integrated photonic devices and circuits have been developed successfully, such as silicon-on-insulator, silicon nitride, and indium phosphide [1–4]. In general, InP platform features the capability of providing a full range of high-performance active photonic devices including laser, amplifier, modulator, and photodetector. However, III-V semiconductors are not the best option for passive photonic devices because the optical waveguides usually have relatively high propagation losses due to the scattering and relatively large bending radii are usually required due to the low index-contrast in the vertical direction [3]. In contrast, silicon photonics has been recognized as one of the most attractive technologies for large-scale photonic integrated circuits (PICs) in the past decade [4–6]. One of the major advantages is that the complementary metal oxide semiconductor-compatibility of the fabrication processes of silicon photonic devices [7,8]. This makes it possible to achieve very large wafer size, very high volume throughput, as well as very low cost [9–13]. As proved by those well-developed foundries, the fabrication of silicon photonic-integrated devices and circuits is robust, high yield, and reproducible. Furthermore, silicon nanophotonic waveguides have ultra-high index-contrast (Δ), enabling ultra-compact photonic devices as well as ultra-dense large-scale PICs [12]. Currently, silicon photonic devices have been developed successfully in terms of performance, size, and reproducibility [11–14]. In this chapter we focus on passive silicon photonic-integrated components and circuits for optical communications.

One might notice that silicon nanophotonic waveguides have some unique mode properties due to the ultra-high Δ and their nano-scale cross section [15–18]. For example, silicon nanophotonic waveguides usually have ultra-high birefringence, very strong-mode hybridness, and large-mode dispersion [19], which are very different from conventional low-Δ optical waveguides (e.g. SiO_2-on-Si buried optical waveguides). As a result, the device structures and the design rules for silicon photonic devices become quite different from the case with conventional low-Δ optical waveguides. In particularity, significant structural asymmetry can be introduced for silicon nanophotonic waveguides and devices. Structural asymmetry has been utilized for the realization of many silicon photonic devices and circuits, e.g. polarization-handling, mode manipulations, as well as power splitting. In Section 29.2, the structures and mode properties of silicon nanophotonic waveguides are discussed. In Section 29.3, a review is given on silicon photonic devices for on-chip polarization-handling, including polarization-beam splitters (PBSs) and polarization splitter-rotators (PSRs). Section 29.4 gives a review for silicon photonic devices for mode-manipulations, including mode converters and mode (de)multiplexers. In Section 29.5, a summary of silicon photonic devices for wavelength-filtering is given, including arrayed-waveguide gratings (AWGs) and micro-ring resonators (MRRs). Finally, a review is given on recent progress of reconfigurable silicon photonic devices and circuits, including Mach-Zehnder interferometer (MZI)/MRR optical switches and reconfigurable optical add-drop multiplexers (ROADMs).

29.2 Silicon Nanophotonic Waveguides

Figure 29.1a shows the cross section of a regular silicon nanophotonic waveguides, which usually has a rectangular silicon core region on a silica insulator layer. The refractive indices of silicon and silica are, respectively, $n_{Si} = 3.48$ and $n_{SiO2} = 1.44$ for the operation wavelength-band around 1550 nm. The upper-cladding is usually made of silica to prefect the silicon core from any damage or contamination. The silica upper-cladding is also needed to isolate the absorption from, for example, metal electrodes, which are used for thermal-tuning. The thickness of the silicon core-layer is popularly chosen as $h_{co} = 220$ nm [4]. In some cases, the thickness of $h_{co} = 340$ nm is also used as a potential option. Nevertheless, the case with $h_{co} = 220$ nm is considered in this chapter if there is no specific statement. As discussed before, silicon nanophotonic waveguides have some unique and interesting mode properties due to the ultra-high Δ and the nano-scale cross section [15–18]. This makes the device structures and the design rules for silicon photonic devices very different from the case with conventional low-Δ optical waveguides.

As shown in Figure 29.1b, which shows the calculated effective indices of the guided-modes in a silicon nanophotonic waveguide with a rectangular silicon core and a SiO_2 upper-cladding, it is not surprise to see that TE- and TM-polarization modes behave very differently. Obviously, the TE_0 mode has much larger effective index than the TM_0 mode because of the stronger mode confinement. Regarding the single-mode condition, one usually chooses the core width around $w_{co} = 0.45$ μm, in which case the birefringence is as huge as ~0.7. The origin of the huge birefringence can be seen from the calculated x-component (E_x) and y-component (E_y) of the electric field for the TE_0 and TM_0 modes when $w_{co} = 0.45$ μm, as shown in Figure 29.2a. As a result, silicon photonic integrated devices usually have giant polarization dependence and thus are designed to work for single polarization only. On the other

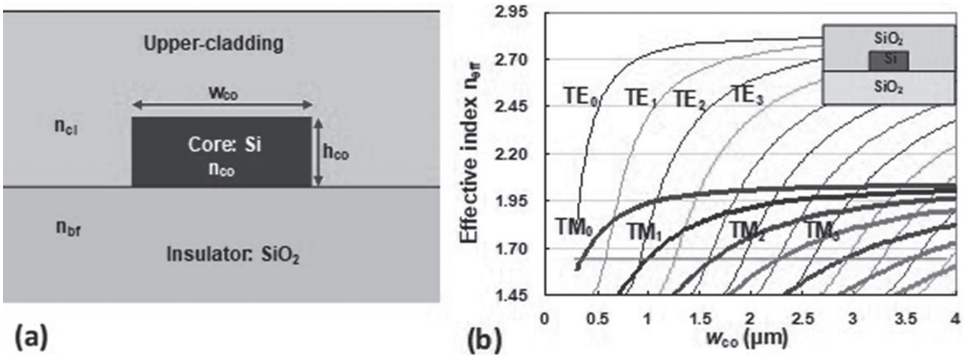


FIGURE 29.1 (a) Calculated effective indices of the guided modes in an SOI strip nanowire with $h_{co} = 220$ nm; (b) mode profiles for the fundamental and higher order modes.

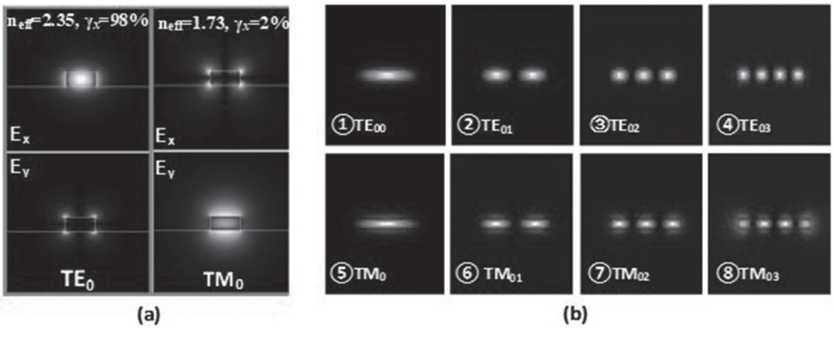


FIGURE 29.2 Calculated mode profiles for silicon nanophotonic waveguides. (a) The TE_0 and TM_0 modes for the case of $w_{co} = 450$ nm; (b) the fundamental and higher order modes for the case of $w_{co} = 450$ nm. Here, $h_{co} = 220$ nm [16]. (Reprinted with permission from Ref. [16], [Optics Express].)

hand, the huge birefringence also provides very excellent opportunity for realizing high-performance on-chip polarization-handling devices, which will be discussed in Section 29.3.

Furthermore, silicon nanophotonic waveguides have very strong mode dispersion. Therefore, it is possible to realize mode-selective manipulation with low excess losses (ELs) and low inter-mode crosstalk for multiple guided-modes supported in a multi-mode silicon nanophotonic waveguide (see Figure 29.2b). This opens the door of utilizing higher order modes in the design of silicon photonic devices by introducing multi-mode optical waveguides. For example, the introduction of higher order modes makes it possible to enable mode-division-multiplexing (MDM), which is a kind of space-division-multiplexed for enhancing the link capacity of data transmissions with multiple mode channels. Furthermore, the higher order modes also become very helpful for realizing some special silicon photonic devices, which cannot be realized with the fundamental mode only. More details will be discussed in Section 29.4.

Besides, as revealed previously [16–20], significant mode hybridness might happen in silicon nanophotonic waveguides when cross-sectional symmetry of a silicon nanophotonic waveguide is broken, which happens very frequently in reality. Figure 29.3a-c show three typical structures for silicon nanophotonic waveguides with cross-sectional asymmetry. The first one is a silicon strip nanowire whose upper-cladding has a different refractive index from the buffer layer [16], the second one is a silicon-strip nanowire with angled sidewalls [17], and the third one is a silicon-ridge nanowire [18].

As an example, the silicon nanophotonic waveguides with an upper-cladding of air and Si_3N_4 are considered and their dispersion curves for the guided modes are shown in Figure 29.4a and b, respectively. From these figures, a significant difference is observed when compared with the result shown in Figure 29.1b. In particularity, there is a gap between the two curves for the TE_1 and TM_0 modes when the core width w_{co} is around 0.7 µm. This happens due to the mode hybridness between the TM_0 and TE_1 modes as discussed previously [16]. More details about the mode hybridness can be found from the calculated polarization ratios γ_x (defined in Ref. [17]) as well as the electrical field profiles given in Figure 29.4d and e, where a silicon strip nanowire with a Si_3N_4 upper-cladding is considered. As it can be seen from Figure 29.4c, the polarization ratios γ_x for modes #2 and #3 have a transition between 100% and 0 around $w_{co0} = 0.73$ µm. This indicates that the mode hybridness happens. For these hybridized modes, the x-component and y-component of the electric fields are even comparable, which is very different from those shown in Figure 29.2a and b. These mode hybridness enables the TM_0-TE_1 mode conversion when using a taper structure whose end-widths are chosen as $w_a < w_H < w_b$ (e.g. $w_H = 0.73$ µm here). This provides a new way for the realization of polarization conversion/rotation if the TM_0-TE_1 mode conversion is very efficient, which is possible when the taper length is sufficiently long, as proposed in Ref. [16].

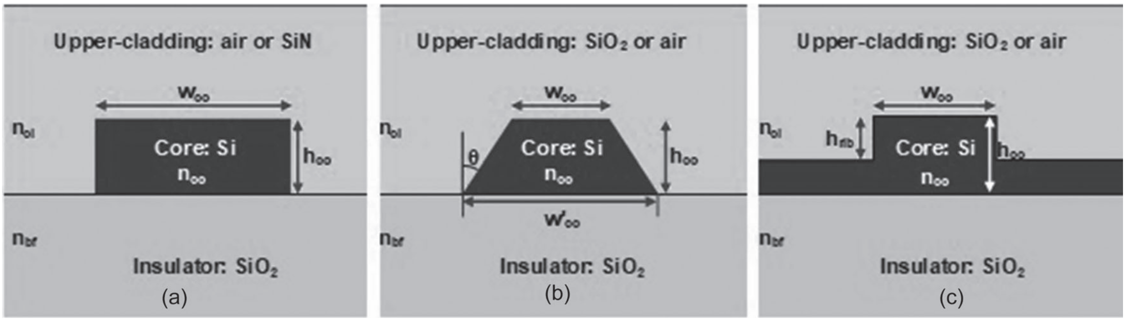


FIGURE 29.3 Three typical silicon nanophotonic waveguides with cross-sectional asymmetry. (a) A silicon strip nanowire with an upper-cladding of air or SiN; (b) a silicon nanowire with angled sidewalls; (c) a silicon ridge nanowire [16]. (Reprinted with permission from Ref. [16] [Optics Express].)

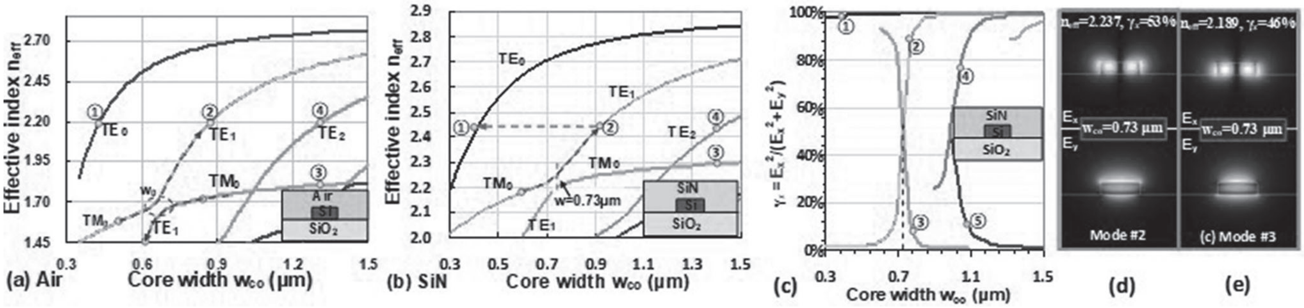


FIGURE 29.4 Calculated effective indices for the guided modes of SOI strip nanowires with an upper-cladding of air (a) and SiN (b) [16]. (c) Calculated polarization ratio $\gamma_x = E_x^2/(E_x^2+E_y^2)$ for a silicon strip nanowire with a Si_3N_4 upper-cladding. Calculated modal field profiles for the hybridized modes when $w_{co} = 0.73$µm: (d) mode #2 and (e) mode #3 [16]. (Reprinted with permission from Ref. [16], [Optics Express].)

29.3 Silicon Photonic Devices for On-Chip Polarization-Handling

On-chip polarization-handling has been playing an important role for many applications, e.g. polarization-transparent PICs [21–23], polarization-division-multiplexing (PDM) coherent optical communications [24–27], and quantum photonics integrated circuits [28–30]. In the past years various structures have been proposed theoretically and demonstrated experimentally [31,32]. However, it is still a challenge to realize ultra-compact on-chip polarization-handling devices with excellent performances, e.g. low EL and high ER (extinction ratio) in a broad wavelength band. Fortunately, silicon nanophotonic waveguides have huge birefringence as high as ~0.7, which makes it possible to realize very excellent on-chip polarization-handling devices, including PBSs and polarization rotators (PRs) [31,32].

29.3.1 Silicon PBSs

As summarized in [32], many silicon PBSs have been presented by using multimode interferometers [33–44], MZIs [45–48], gratings [49], photonic-crystal (PhC) structures [50–53], mode-evolution structures [54], AWGs [55], MRRs [56], directional couplers (DCs) [57–60], and asymmetric DCs (ADCs) [61–80], etc. Some special topological structures were recently introduced to realize extremely small PBSs (e.g. ~2.4 × 2.4 μm^2) [81]. Among these structures, ADCs have been well recognized as one of the most attractive options for realizing PBSs [31,32] because of the performance excellence and design simplicity. An ADC can be constructed by using two or more different optical waveguides in the coupling region, and currently compact and broadband ADC-based PBSs with very excellent performances have been demonstrated [82].

When using ADCs for realizing PBSs, the basic idea is to manipulate the phase-matching condition of both polarizations [32]. One should choose the structural parameters optimally, so that the phase-matching condition is satisfied for one polarization and unsatisfied for the other polarization. This is possible for an ADC which consists of two optical waveguides with *different birefringence*. In order to achieve a compact ADC, it is preferred to make the phase-matching condition satisfied for the polarization with weak mode-confinement, and consequently a complete cross-coupling can be achieved within a short coupling region. In contrast, for the other polarization with strong-mode-confinement, the phase-matching condition is broken *automatically* due to the birefringence. According to the coupled-mode theory, the cross-coupling can be depressed significantly if the phase-mismatching is huge. It can be seen that the design of such PBSs is very convenient because one needs to optimize the waveguide dimensions for one polarization only according to the phase-matching condition.

In order to construct an ADC, one of the easiest options is to combine different types of optical waveguides. For example, one can use the combination of a regular silicon strip nanowire and a special optical waveguide, e.g. silicon nano-slot waveguides [62–65], sub-wavelength-grating (SWG) waveguides [83,84], silicon hybrid plasmonics waveguides (HPWGs) [71–73], etc. When using nano-slot waveguides proposed in 2004 [85], the PBS can be realized with a very compact footprint, e.g. 6.9–16 μm [26,62–65]. When introducing an SWG waveguide [86], whose effective index and birefringence can be modified flexibly by adjusting the duty circle [86], the design of polarization-handling devices is very convenient [83,87,88] even when the regular silicon photonic waveguide does not have a huge birefringence (e.g., h_{co} = 340 nm) [84]. On the other hand, one should notice that the fabrication for nano-slots and SWGs is not easy because of the deep sub-wavelength feature size (~100 nm). As an alternative, a silicon HPWG [89] is attractive for realizing ultra-compact PBSs [71–73]. For example, the total size of the designed PBS in [73] is as small as ~1.9 × 3.7 μm^2. In this case, the problem is that the metal absorption introduces some EL (which is usually acceptable).

As a result, an ADC without any special waveguide structures is still preferred regarding the fabrication convenience and the device performance. The novel ADCs with a multi-mode middle optical waveguide [74] and a bent DC [61–66] provide an attractive option for realizing PBSs. For the ADC with a multi-mode middle optical waveguide [74], the waveguide widths are designed to achieve complete evanescent coupling for one polarization between the fundamental mode in the narrow optical waveguide and the higher order mode in the wide optical waveguide. Meanwhile, the evanescent coupling for the other polarization is very weak due to the large phase mismatch. Recently such an ADC was extended further for realizing high-ER polarizers [90].

The PBS based on bent DCs was proposed for the first time in 2011 [61] and was realized experimentally in 2013 [66]. This idea has become one of the most popular designs [91]. For the design of bent DCs [61,92], one can manipulate the phase-matching condition by modifying the core widths as well as the bending radii of the two bent waveguides. This idea has also been extended to a triple-waveguide bent DC recently [76–78], for which the measured ERs are >20 dB for TE polarization, and >15 dB for TM polarization in a broad bandwidth of 90 nm. In order to improve the performance of the PBSs based on bent DCs, an effective approach is introducing on-chip polarizers [75–79,82]. There are many compact on-chip polarizers available, which were realized by utilizing, for example, an SWG waveguide [93], a hyper-uniform disordered network waveguide [94], a silicon HPWG [95], etc. In [75] a compact PBS consisting of a bent DC cascaded with a SWG TM-passed polarizer was proposed and realized with a low EL and a high ER (>25 dB) over a ~30nm bandwidth. In [96], a ~26 μm-long PBS was demonstrated by using cascaded triple-bent-waveguide DCs, which has an ER of ~30 dB and a low EL of 0.3~0.8 dB in a ~50 nm bandwidth. More recently the cascaded structure of bent DCs was proposed for realizing a near-perfect PBS [82], as shown in Figure 29.5a. Here DCs #1 and #2 are designed according to the phase-matching condition to achieve efficient cross-coupling for TM polarization. DC #3 was introduced to filter out the residual power of TM polarization at the through port, in which way one can greatly improve the bandwidth with a high ER for TM polarization. For TE polarization, DC #2 helps greatly improve the ER by filtering out the undesired weak cross-coupled power in DC #1. As shown by the measurement results

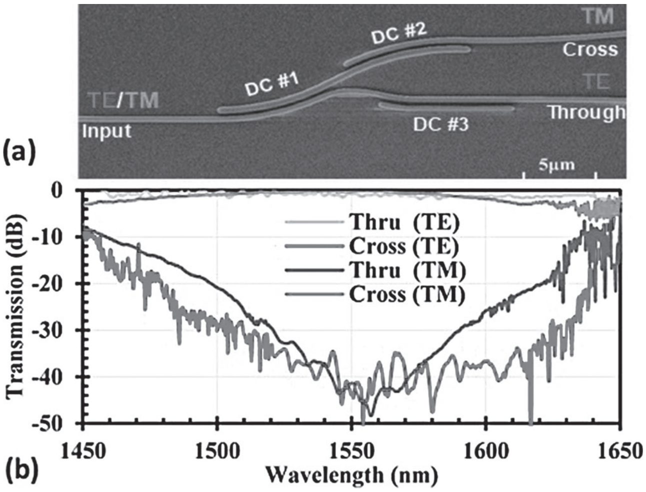


FIGURE 29.5 The PBS based on cascaded bent DCs: (a) the fabricated device; (b) the measurement results [82]. (Reprinted with permission from Ref. [82], [Optics Express].)

given in Figure 29.5b, it can be seen that the bandwidths are as large as ~126 nm, ~97 nm, and ~70 nm for achieving an ER of >20 dB, >25 dB, and >30 dB, respectively. Meanwhile, the measured ELs are <1 dB and <0.5 dB over a bandwidth of ~137 nm and ~91 nm, respectively. In addition, this PBS has a large fabrication tolerance (e.g., ±20 nm) for the variation of the core-widths, which makes it very attractive for many silicon PICs.

29.3.2 Silicon PRs

The realization of on-chip PRs is often challenging as a planar optical waveguide usually has a very good ability of polarization maintaining. Generally speaking, it is necessary to enable the mode hybridness between the two orthogonal polarizations, which can be achieved possibly by introducing some special asymmetric structures. Then polarization rotation can be realized by using the hybridized-mode *evolution* as well as the hybridized-mode *interference*. In the past decade great efforts have been made and on-chip PRs have been realized by using asymmetric waveguide structures with, for example, angled sidewalls, bends, corner-cut cross sections, etc. [97–108]. There are mainly five types of structures, as summarized in Figure 29.6a–r [19].

Type I is based on the structures with a uniform straight polarization rotation section (see Figure 29.6a–e), and the hybridized-mode interference is utilized [109–114]. Here the mode hybridness is generated by using an asymmetric waveguide structure with, for example, a cut-corner (Figure 29.6a) [109,110], dual-trenches (Figure 29.6b) [111], a poly-silicon strip (Figure 29.6c) [112], a SiN square (Figure 29.6d) [113], as well as a metal strip (Figure 29.6e) [114]. In this case, the geometry of the waveguide cross section should be designed optimally to achieve a polarization ratio of $\gamma_x \approx 50\%$, which is corresponding to an optical axis of ~45°. As a result, two hybridized-modes can be excited with equal mode-excitation ratios and two-mode interference happens in the polarization-rotation section. When its length is chosen optimally according to the two-mode interference theory, polarization rotation is achieved within a length as short as several micrometres [110]. Similarly, the approaches of introducing other materials (e.g. SiN and metal) also provide good options regarding the performance and the footprints of the devices, as shown in Figure 29.6c–e [112–114]. For these designs, the width and the etching-depth of the cut-corner should be well controlled. In [111], a silicon nanowire with dual subwavelength trenches was proposed and only a single-etching process is needed by smartly taking advantage of the reactive-ion etching lag effect (Figure 29.6b). The problem is that the sub-wavelength trench is as narrow as several tens of nanometres (e.g., 60 nm), in which case a regular deep-UV lithography does not work.

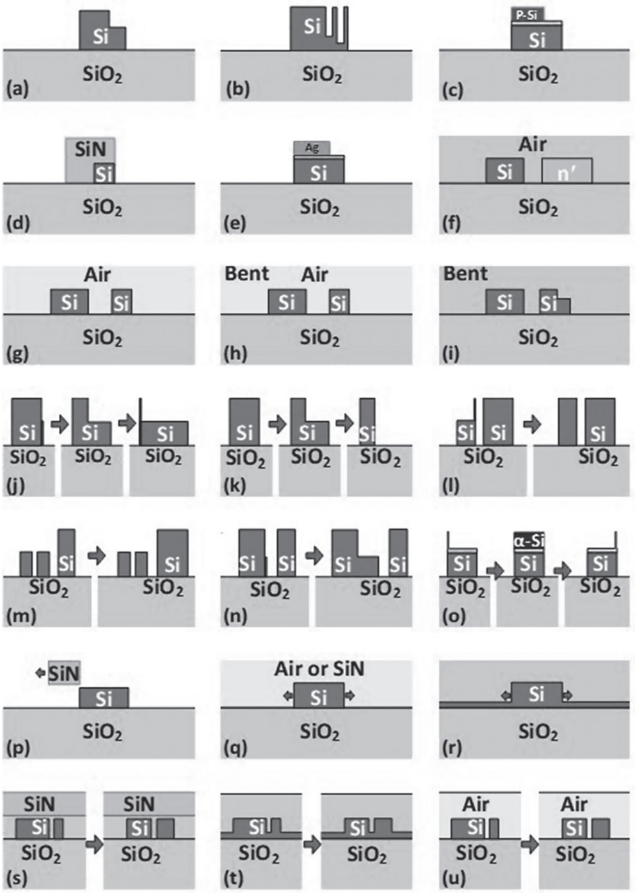


FIGURE 29.6 Cross sections of the waveguide structures for PRs (or PSR), including five types of structures [19] (Copyright in 2018, IEEE). Type I: the structures with a uniform straight polarization rotation section, as shown in (a) [109], [110], (b) [111], (c) [112], (d) [113], and (e) [114]; Type II: the structures with ADCs, as shown in (f) [83], (g) [115], (h) [98], and (i) [116]; Type III: the structures with double-etched adiabatic tapers, as shown in (j) [117], (k) [118], (l) [119], (m) [120], and (n) [121]; Type IV: the structures based adiabatic tapers with other materials, as shown in (o) [122], and (p) [123]; Type V: the structures based single-etched adiabatic tapers with single or dual cores, as shown in (q) [16,125,114], and (r) [18,124], (s–t) [126], and (u) [124].

As an alternative, Type II is based on the structures with ADCs (see Figure 29.6f–i) [83,98,115,116]. The mechanism is utilizing the polarization coupling in an ADC consisting of an optical waveguide with mode hybridness, enabled by the introduced air-cladding [83,98,115] or cut-corner [116]. For example, in [83] the proposed polarization splitter-rotator (PSR) was designed with an ADC consisting of a silicon strip nanowire and an SWG waveguide (Figure 29.6f). It is shown numerically that the designed PSR has a low polarization-conversion loss of ~0.4 dB over the C-band. Unfortunately, it is still sensitive to the variation of the SWG duty-cycle. In [98], a PSR was realized with a compact footprint (8.7 μm), a low loss (<1 dB), and a decent crosstalk (<–18 dB) by using an air-cladded bent DC (Figure 29.6h). More recently, in [116] a PSR based on a bent DC with a corner-cut silicon nanowire was proposed (Figure 29.6i), and the TM-TE polarization conversion loss is ~0.1 dB and the ER is ~27 dB in theory.

It is also possible to realize PSR by using the structures based on a double-etched adiabatic taper with cross-sectional asymmetry (Type III), as shown in Figure 29.6j–n [117–121]. For Type III, a regular double-etching process is needed for this type while the fabrication is relatively easy because of the large fabrication tolerance for long adiabatic tapers. For Type IV, the PSR is realized with the structures based on adiabatic tapers with an additional layer of other materials (Figure 29.6o and p), e.g. α-Si [122] and SiN [123], which definitely requires additional fabrication processes. For Type V, the PRs are realized by using a single-etched adiabatic taper based on a silicon nanowire with an air (or SiN) upper-cladding [16,124–126], or silicon ridge waveguides [18,124,126], as shown in Figure 29.6q and r. It is also possible to use the silicon nanowire with dual cores (Figure 29.6s–u) [124,126], in which case the evolution of the hybridized super-mode in the dual-core taper is considered. This novel concept for realizing PSRs with the TM_0-TE_1 mode conversion in an adiabatic taper on silicon was proposed in 2011 [16], as shown in Figure 29.7a. For the adiabatic taper section, the core-widths are chosen optimally to enable mode hybridness while the length is long enough to be adiabatic, so that the TM_0-TE_1 mode conversion can be very efficient [16]. In particular, a three-segment taper was proposed in [16] in order to make the adiabatic taper compact (see Figure 29.7a). The cascaded ADC is designed according to the phase-matching condition to make the TE_1 mode in the wide waveguide coupled efficiently to the TE_0 mode of the adjacent narrow waveguide. Finally, the launched TM_0 mode is converted into the TE_0 mode at the cross port, as shown in Figure 29.7b. For the launched TE-polarization mode, no inter-mode conversion happens in the adiabatic taper and almost no cross-coupling happens in the ADC due to significant phase-mismatching (see Figure D6.7c). Such a PSR has a broad bandwidth of >70 nm for achieving an ER of >10 dB, and it is possible to achieve a larger bandwidth and higher ER by choosing a longer taper [16]. The ER of the PSR can be improved further by introducing a multimode interference (MMI)-mode filter cascaded at the through port (see Figure 29.8a) [127]. Figure 29.8b and c show the measured spectral responses. It can be seen that the fabricated PSR has excellent performances with a low EL (~0.5 dB) and a high ER (~20 dB) in the wavelength range of 1530~1580 nm, which satisfies the requirements for most applications with dual polarizations. Numerous similar designs with modified structures have been proposed in the recent years [124–133]. Similar polarization mode conversion was also observed in silicon ridge nanowires [15], and silicon nanowires with angled sidewalls [17]. This idea can also be extended for realizing polarization splitting-rotation for other wavelength bands [131,132].

29.4 Multi-mode Silicon Photonics

For most PICs, higher order modes are often not allowed due to the unacceptable crosstalk and loss caused by the MMI [11,134]. As a result, the design usually obeys the single-mode condition and only the fundamental mode is involved. As mentioned in Section 29.2, currently the situation is changing

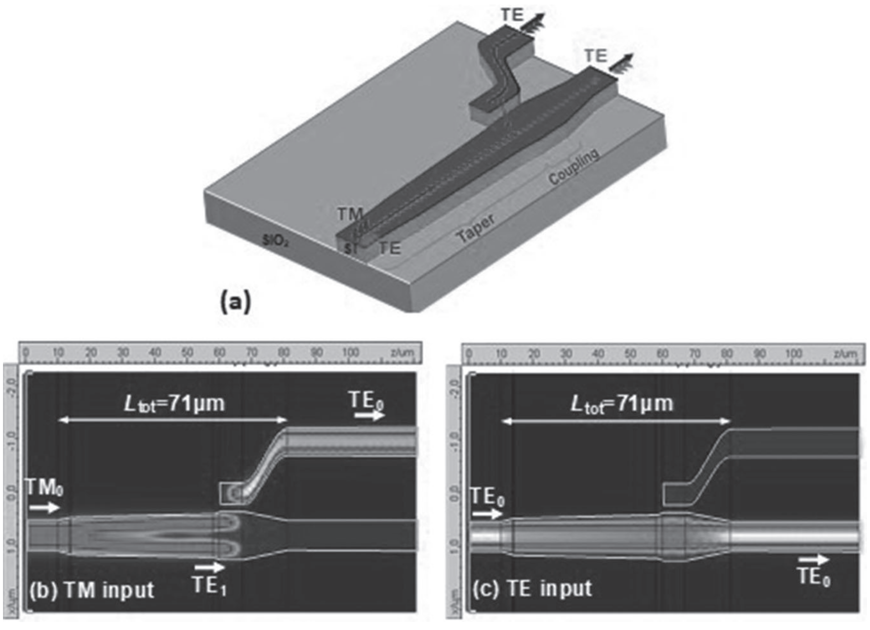


FIGURE 29.7 A PSR consisting of an adiabatic taper and an ADC: (a) the configuration; simulated light propagation in the designed PSR when the TE_0 (b) and TM_0 (c) modes are launched [16]. (Reprinted with permission from Ref. [16], [Optics Express].)

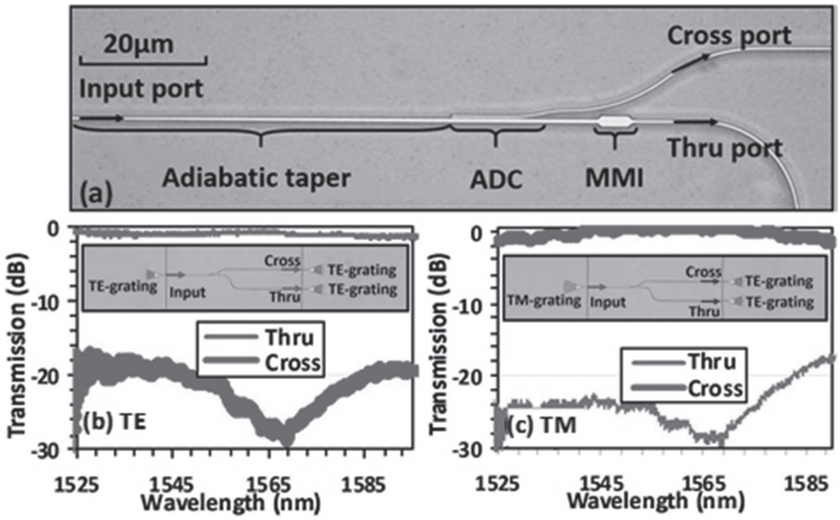


FIGURE 29.8 (a) An improved PSR with an MMI mode filter [127]; measured results for TE (b) and TM (c) polarization modes. (Reprinted with permission from Ref. [127], [Optics Letter].)

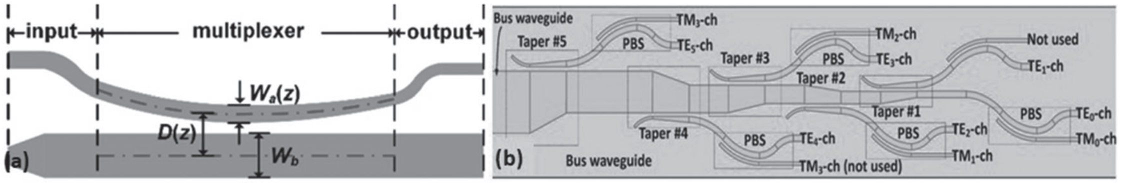


FIGURE 29.9 (a) An STA taper for mode (de)multiplexers [167]. (Reprinted with permission from Ref. [167], [Optics Express]); (b) a mode (de)multiplexer with dual polarizations using cascaded dual-core adiabatic tapers [168]. (Copyright in 2018, Laser Photonics Review.)

because silicon nanophotonic waveguides have very strong-mode dispersion (see Figure 29.1a) [12,19], in which case mode-selective manipulation becomes possible. This opens the door of developing multi-mode silicon photonics, for which multimode waveguides supporting higher order modes are introduced in the device design. As one might notice, currently multi-mode silicon photonics is becoming very attractive because of the following two major reasons [135]. One is that the introduction of higher order modes makes it possible to increase the channel number for data transmission in MDM systems. The other one is that it can improve the flexibility of device designs [136,137]. In recent years, numerous multi-mode silicon photonics have been developed successfully for MDM systems, including on-chip multi-channel mode (de)multiplexers [138–173], multi-mode waveguide bends [174–184], multi-mode waveguide crossings [185–189], reconfigurable multi-mode silicon PICs [190–193], multi-mode chip-fibre couplers [194–199], etc. In addition, multi-mode waveguides with higher order modes are also introduced for the realization of some special silicon photonic devices, which cannot be realized with the fundamental mode only. For example, high-performance on-chip polarization-handling devices with higher order modes [74,90] and on-chip add-drop optical filters based on multi-mode Bragg gratings [200,201] have been demonstrated. Some emerging applications with higher order modes have also been reported recently [29,202]. The details about multi-mode silicon photonics can be found in the recent comprehensive review paper [135,203]. Here we focus on the parts of mode (de)multiplexers and sharp multimode waveguide bends regarding that they are key elements for multi-mode silicon photonics.

29.4.1 On-Chip Mode (De)Multiplexers

As one of the most important representative elements for MDM systems, mode (de)multiplexers are used to combine/separate the mode channels with low losses and low crosstalk. In the past years, various mode (de)multiplexers on silicon have been developed by using MMI couplers [139,140], topology structures [163], staged couplers [141], adiabatic mode-evolution couplers [142–144], and ADCs [145–149]. Among them, the design of the mode (de)multiplexer based on MMI couplers is relatively complicated [139,140], and it is inflexible for achieving more mode channels. The design with topology structures is becoming attractive for mode (de)multiplexers regarding that the potential for shrinking the footprints (e.g. 2.6 μm × 4.2 μm in [163]). However, the fabrication is still not easy due to the small feature size. Furthermore, it is still a big challenge for realizing high-performance mode (de)multiplexer with multiple channels.

In contrast, the structure of adiabatic mode-evolution couplers provides an excellent option because of the high flexibility for more mode channels. Furthermore, one can achieve ultra-broadband mode (de)multiplexers when the length of the coupling region is sufficiently long [142–144]. For example, multi-channel mode (de)multiplexers can be realized with asymmetric Y-junctions cascaded in parallel [137] as well as in series [157]. However, the fabrication is still difficult due to the tiny gaps in the Y-junctions. As an alternative, adiabatic dual-core tapers in cascade have been used to enable multi-channel mode-multiplexing [161–168], and one can choose sufficiently large gaps between the dual cores for easy fabrication [150]. For the design of the adiabatic tapers, the scheme of shortcuts to adiabaticity (STA) was proposed recently [163–167]. Figure 29.9a shows an example of a four-channel mode (de) multiplexer with cascade STA tapers [167], in which case the bus waveguide has a uniform core width while the gap width varies gradually along the propagation direction. The realized four-channel mode (de)multiplexer has a <1.3 dB EL for the conversion of TE_0-TE_i and <–24 dB inter-mode crosstalk in the wavelength-band ranging from 1500 nm to 1600 nm. More recently, a 10-channel mode (de)multiplexer with dual polarizations was proposed and realized with cascaded dual-core adiabatic tapers [168], as shown in Figure 29.9b. In this design, the desired highest order modes of TE- and TM polarizations can be extracted simultaneously for each stage. These two extracted mode-channels are then separated by using a PBS based on bent DCs. For the fabricated 10-channel mode (de) multiplexer, all TM- and TE-mode channels have low crosstalk (–15 ~ –25 dB) as well as low ELs (0.2~1.8 dB) over a broad wavelength band of ~90 nm. This makes it WDM-compatible and thus suitable for high capacity on-chip optical interconnects (Figure 29.10).

Another popular structure for mode (de)multiplexers is cascaded ADCs because of the structural simplicity, the scalability, and the flexibility for mode-selective coupling. In 2012, a two-channel mode (de)multiplexer was demonstrated with an ADC based on SiO_2 optical waveguides [151], and the first four-channel mode (de)multiplexer on silicon was proposed with multiple ADCs in cascade [145,146]. In this design, the bus waveguide consists of multiple straight segments with different core-widths, while the width of the access waveguide is fixed to be ~400 nm. The width of any straight segment in the bus waveguide is chosen according to the phase-matching condition [146]. In order to connect these straight segments with different core widths, some adiabatic taper sections are introduced, which also helps greatly filter out the undesired residual power of higher order modes when de-multiplexing. As a result, broadband multi-channel mode (de)multiplexers can be realized with low loss (<1dB) and low crosstalk (<–20dB) [146], as shown in Figure 29.11. Grating-assist ADCs [147] and a micro-ring-coupled ADCs [148] have also been realized successfully for narrow-band multi-channel mode multiplexer. It is also possible to use tapered ADCs so that the fabrication tolerance can be improved, as demonstrated in [149]. These on-chip mode (de)multiplexers can be integrated with WDM/PDM devices to enable hybrid WDM-MDM multiplexing technology [203–214].

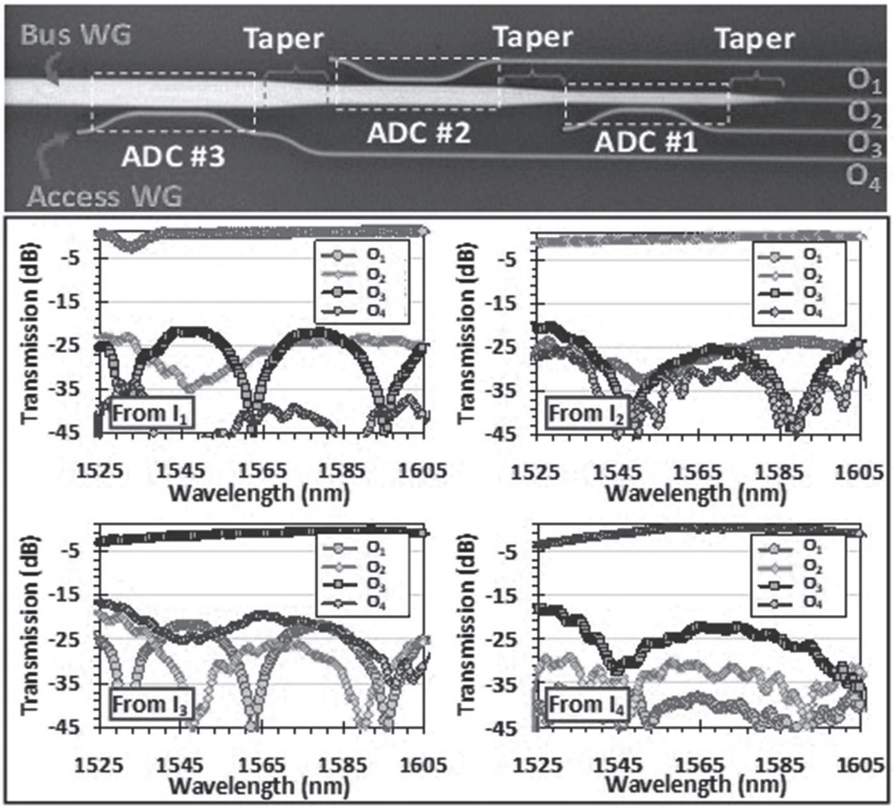


FIGURE 29.10 A four-channel mode de-multiplexers based on cascaded ADCs [146] and the measured transmissions at output ports (O_1, O_2, O_3, and O_4) of a mode de-multiplexer when light is input from port I_1~I_4 of an mode multiplexer on the same chip. (Reprinted with permission from Ref. [146], [Optics Letter].)

29.4.2 Sharp Multi-mode Waveguide Bends

As it is well known, waveguide bends are usually inevitable for building any PICs. In particular, sharp waveguide bends are often desired to achieve compact footprints, which is very easy for single-mode silicon nanophotonic waveguides owing to the ultra-high index contrast [215–220]. However, it is a big challenge for the realization of sharp *multimode* waveguide bends because significant mode-mismatching loss as well as inter-mode crosstalk might be introduced [174], as shown in Figure 29.11a–d. As a result, when using a regular arc-bend for a wide multi-mode waveguide (e.g. w_{co} = 4 μm in [174]), the bending radius should be chosen as large as, for example, millimetres. A potential approach to achieve sharp multimode waveguide bends is using a kind of vertical multimode waveguide supporting multiple modes in the vertical direction [175]. However, this does not work for the regular case with a thin top-silicon layer, e.g. h_{co} = 220 nm, which has been used popularly. In order to solve this problem, one should introduce some special designs to minimize the mode mismatch between the straight section and the bent section.

For example, when making the *curvature* of the bent section varies from zero to a given value gradually, it is possible to remove the mode mismatch at the straight-bent junction and all guided-modes can propagate in the bent section with negligible mode distortion [176]. This idea was used for single-mode waveguide bends previously in order to reduce the mode-mismatch loss [217], and some impressive results were also reported for multimode waveguide bends recently [176,177]. In [176], a 90°-bend with Bezier curve was designed to enable low-crosstalk (–23 dB) propagation of three TE-polarization modes. In [177], a multi-mode waveguide bend with modified Euler curve was realized. Figure 29.12a–d shows the simulated light propagation in the designed Euler bend with an effective radius R_{eff} = 45 μm. As it can be seen, almost no MMI happens. In contrast, serious MMI is observed for a regular arc-bend with R = 45 μm, as shown in Figure D6.311a–d. For the fabricated Euler S-bend, the mode-mismatching loss is <0.5 dB and the inter-mode crosstalk is <–20 dB over a broad wavelength band from 1520 to 1610 nm for all four TM-polarization mode-channels [177]. This design can be extended for more channels of dual polarizations flexibly. For example, low-loss and low-crosstalk transmission of ten mode-channels was demonstrated in a 2.3 μm-wide bus waveguide with sharp waveguide bend with R_{eff} = 40 μm [1].

In order to further reduce the footprint of the waveguide bends, another approach is to utilize some special *mode converter* between the straight and bent sections, in which way their mode-mismatch can be reduced. For example, in [178,179] a 5-μm-long mode converter was proposed and demonstrated for a 1-μm-wide multi-mode waveguide bend supporting the TE_0 and TE_1 modes. With this design, the inter-mode crosstalk for the TE_0 and TE_1 modes is <–22 dB in a 100 nm bandwidth even when R = 5 μm (i.e., R_{eff} = 10 μm). In [180] a sharp multi-mode waveguide bend with R_{eff} = 3 μm was realized for two guided modes by combining two single-mode arc-bends with a phase

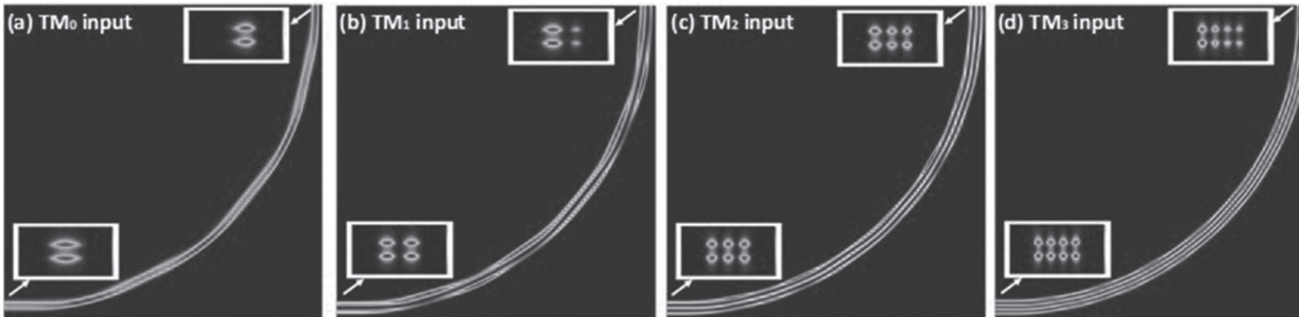


FIGURE 29.11 (a–d) Simulated light propagation in a regular arc-bend with R = 45 μm when the TM_i mode of a straight waveguide is launched [177]. (Reprinted with permission from Ref. [177], [Optics Express].)

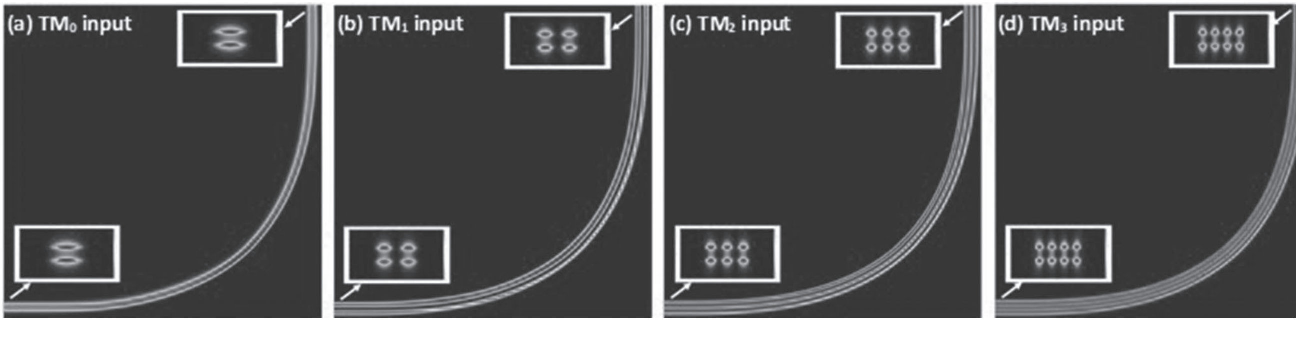


FIGURE 29.12 Simulated light propagation in the designed Euler-bend with R_{eff} = 45 μm (a–d) when the TM_i mode of a straight waveguide is launched [177]. (Reprinted with permission from Ref. [177], [Optics Express].)

difference of 4π and mode converters based on Y-junctions. The EL is 0.8 dB and the inter-mode crosstalk is –20 dB in theory. Unfortunately, it is not easy to extend these designs for the case with more than two modes [178–180]. More recently, a novel adiabatic mode converter with a metamaterial-assistant PMMA upper-cladding was introduced for realizing sharp multi-mode waveguide bends [181]. As demonstrated experimentally [181], all four TM mode channels have ELs of <1 dB and inter-mode crosstalk of <–20 dB over a wavelength band of 80 nm when choosing $R = 30\,\mu\text{m}$ and a mode converter as long as $L_{\text{tp}} = 15.8\,\mu\text{m}$.

In order to realize ultra-sharp multi-mode waveguide bends, it is also possible to minimize the straight-bent mode mismatching by modifying the effective refractive index profile in the bent section. In [174], a multi-mode waveguide bend with optimally gradient core heights was proposed and designed according to the transformation optics theory. As demonstrated, the fabricated three-mode waveguide bend has an EL of −2.6 dB when $R = 78.8\,\mu\text{m}$. However, one should notice that the fabrication is quite complicated since a grayscale-lithography process is needed. Alternatively, one can also modify the effective index profile in the bent section as desired by introducing sub-wavelength structures (e.g., nano-holes as well as nano-slots [182,183]), which was utilized for the design of single-mode waveguide bends previously [221] and recently has been extended for multi-mode waveguide bends [182,183]. In [182], a compact multi-mode bend was demonstrated by using a pair of Y-junctions with the assistance of nano-holes. Similarly, a sharp waveguide bend with many nano-holes was demonstrated for the fundamental mode in [183] (see Figure 29.13a). More recently, a sharp multi-mode waveguide bends with shallowly etched SWG structures was proposed and demonstrated, as shown in Figure 29.13b [184]. In this design, the introduced SWG structure is asymmetric spatially so that the equivalent refractive index profile of the bent section matches well with that of the straight section. For the realized multi-mode waveguide bend, the EL is as low as 0.5~0.8 dB and the crosstalk is <–20 dB for all three TE modes even for a very sharp bend with $R = 10\,\mu\text{m}$.

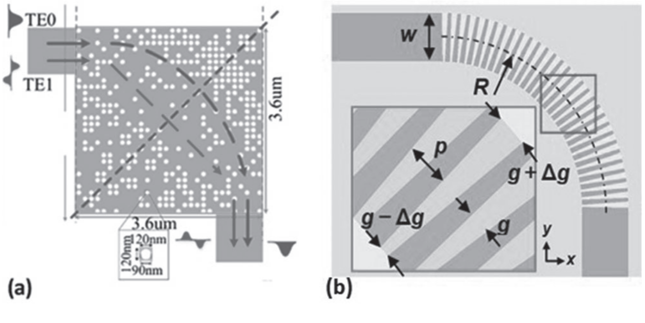


FIGURE 29.13 (a) A two-mode waveguide bend consisting of a pair of equivalent Y-junctions designed with nano-holes [182]. (Reprinted with permission from Ref [182], [Optical Fiber Communication Conference]); (b) a multi-mode waveguide bend with sub-wavelength structures [184]. (Copyright in 2019, Laser Photonics Review.)

29.5 Wavelength-Selective Silicon Photonic Devices

For WDM systems with many wavelength-channels, wavelength-selective photonic devices have been always playing very important roles. One of the fundamental principles for the realization of wavelength-selective devices is utilizing the interference of two or multiple optical beams. As one of the most representative two-beam interferometers, MZIs have been used very widely. In order to achieve a flat-top spectral response, a promising approach is using the design with multiple cascaded MZIs, for which coupling coefficients as well as the critical phase delays of the stages should be optimized carefully [222]. Such a cascaded MZI filter with flattened spectral responses is useful for two-channel optical inter-leavers [223–226] and can also be used for realizing multi-channel CWDM (de)multiplexers with a large channel-spacing (e.g. 20 nm) [227].

For DWDM systems, multi-beam interferometers are preferred because it enables narrow-passband wavelength filtering over a large wavelength period. The first type of multi-beam interferometers is AWGs [228–246] and echelle diffraction gratings (EDG) [247–251]. These two are quite similar, while AWGs are more popular as DWDM devices than EDGs because of the stronger dispersion ability. Therefore, the details about EDGs will not be included in this chapter. The second type of multi-beam interferometer is based on optical cavities, e.g. F-P cavities, MRRs, etc. [252–255]. One might notice that on-chip F-P cavities are not popular for DWDM systems. The reason is that the Q-factor is usually low and a circulator is often needed for separating the input and the reflected beams. In contrast, MRRs are much more popular because four ports are available (i.e. input, add, drop, through) and sufficiently high Q-factors can be achieved. It is very convenient to realize an array with $N \times N$ elements to deal with N wavelength channels by cascading the MRRs. More details about MRRs will be given below. The third type of multi-beam interferometer is waveguide Bragg gratings. However, conventional Bragg gratings also need to work together with circulators, which is similar to regular F-P cavities. Fortunately, this issue has been solved by using the multi-mode waveguide gratings developed recently, which will be discussed below.

29.5.1 AWG-Based Optical Filters

In the early 1990s, Smit et al. proposed AWGs [256], which was further developed by Dragone (using star couplers) [257], and Takahashi et al. (using external lenses) [258]. An AWG usually consists of input waveguides, arrayed waveguides, output waveguides as well as one or two free propagation regions [228]. An AWG (de)multiplexer usually has one or N_{i} input waveguides, N_{a} arrayed waveguides as well as N_{o} output waveguides, as shown in Figure 29.14a. In particular, the adjacent arrayed waveguides have a constant length difference ΔL, which is an important parameter to influence the dispersion ability of an AWG [229]. When used for DWDM systems with very narrow channel spacing, one usually has to choose a large ΔL and the AWG size becomes very large. For those AWGs commercialized based on low-Δ optical waveguides (e.g. SiO_2-on-Si buried waveguides [230,231]), the

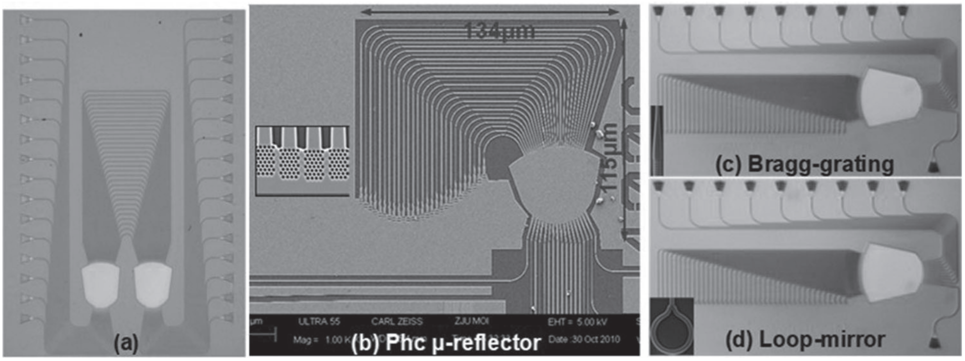


FIGURE 29.14 (a) An *N* × *N* AWG (de)multiplexer on silicon; reflective AWG (de)multiplexer integrated with (b) PhC μ-reflectors [242]; (c) Bragg grating μ-reflectors; (d) loop-mirror μ-reflectors. (Reprinted with permission from Ref. [242], [Optics Letter].)

device size is usually as large as several cm^2 because large bending radii are required for low bending losses. In contrast, when using silicon nanophotonic waveguides, one can realize AWG (de)multiplexers as small as ~0.01 mm^2 [232–240]. In order to reduce the AWG's footprint, novel layer designs for AWGs have been introduced [241]. Among them, reflective AWG de-multiplexer integrated with micro-reflectors at the end of each arrayed waveguide is attractive because the AWG size is reduced by half and the layout design of AWGs also becomes flexible [242]. As shown in Figure 29.14b–d, the AWGs with three types of micro-reflectors based on PhC structures, Bragg gratings as well as loop mirrors have been demonstrated. The former two micro-reflectors are ultra-compact but polarization-sensitive, while loop mirrors can work for two polarizations if the bending radius is large enough.

In order to realize high-performance silicon AWGs/EDGs with ultra-small footprints, great efforts have been made to improve the fabrication processes as well as the structure designs in the past ten years. For example, the on-chip loss of AWGs is lowered to around 2 dB by introducing some specific designs, e.g. bi-level taper structures for AWGs [11] and Bragg-grating reflectors for EDGs [248]. The crosstalk of silicon-nanowire AWGs has also been reduced significantly to −20~−25 dB, which was realized by optimizing the waveguide layout or choosing wide cores in the straight regions to minimize the phase errors due to the random core-width variation [236–240]. In order to improve the channel uniformity of AWGs, one can introduce some auxiliary waveguides between adjacent arrayed waveguides [243]. In particular, special dual-tapered auxiliary waveguides were proposed for silicon AWGs regarding the strong confinement and the sub-micron size of silicon nanophotonic waveguides for the first time [243]. Recently, this idea was applied to realize cyclic AWGRs on silicon [244]. The demonstrated AWGRs show good channel-uniformity (<0.5 dB), which is significantly improved by at least 1.5 dB compared with conventional AWGRs.

29.5.2 MRR-Based Optical Filters

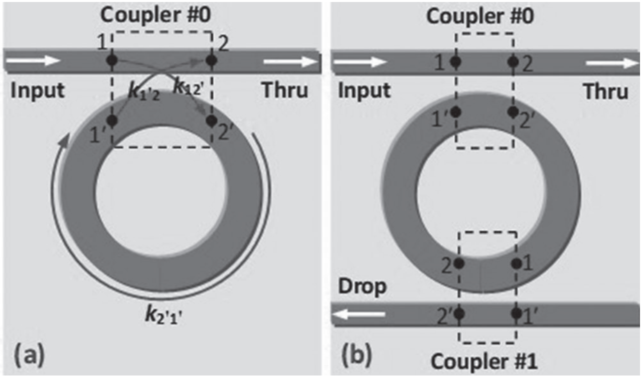


FIGURE 29.15 Schematic configurations of MRR devices: (a) an all-passed optical filter; (b) an add-drop optical filter.

An MRR usually has one or two 2 × 2 couplers to be an all-passed filter or add-drop filter, as shown in Figure 29.15a and b, respectively. For an MRR-based optical filter, the important parameters include the resonance wavelength λ_{res}, the free spectral range (FSR), the 3dB-bandwidth $\Delta\lambda_{3\text{dB}}$, and the ER [252]. The bandwidth $\Delta\lambda_{3\text{dB}}$ and the ER are determined by the cavity loss and the coupling ratios of the couplers. When a small bandwidth $\Delta\lambda_{3\text{dB}}$ (i.e. a high Q factor) is desired, one needs to reduce the coupling ratio as well as the micro-ring waveguide loss (including the scattering loss and the bending loss) [259]. In order to have a high ER, the coupling ratio should be chosen appropriately according to the critical coupling condition [252]. As an important role for MRRs, 2 × 2 couplers can be realized by many structures, including straight DCs [260], MMI couplers [261], Mach-Zehnder couplers [262] as well as bent DCs [263,264], etc. Among them, bent DCs have been recognized as a very excellent option because of the following advantages compared to straight DCs and MMI couplers. First, the coupling ratio can be chosen very flexibly and can be sufficiently high when the waveguide width as well as the angle of the bent coupling region are chosen appropriately. Second, there is almost no EL in theory. Third, the cavity length (=2π*R*) is as short as a standard micro-ring and thus the FSR of MRRs can be maximized to the limit determined by the bending loss. In the following parts, recent progresses on MRRs with bent DCs will be presented, including MRRs with ultra-large FSRs, box-like spectral responses as well as high polarization-selectivity.

29.5.2.1 MRR-Based Optical Filters with Large FSRs

Regarding that a large wavelength window might be required for WDM systems, it is desired to have MRR filters with an ultra-large FSR and a large tunability range of wavelength [265]. For a single-ring resonator, the FSR is inversely proportional to the cavity length (i.e. the bending radius) [266,267] and thus smaller MRR is desired to achieve a larger FSR. A straightforward way to increase the FSR is to reduce the bending radius. Definitely, the minimal bending radius is limited by the allowable bending loss. In the past years, great efforts have been made to achieve MRRs with ultra-large FSRs [266–268], as summarized in Table 29.1. For example, in 2006, an add-drop optical filter with a large FSR of 47 nm and a 3 dB-bandwidth of ~0.24 nm was realized with a SiO_2-cladded MRR with $R = 2\,\mu m$ [269]. An all-pass MRR with $R = 1.5\,\mu m$ was also reported with an FSR as large as 62.5 nm and a 3 dB-bandwidth of 0.17 nm by using a 450 nm × 250 nm single-mode micro-ring waveguide [266]. The bending radius was even reduced to $R = 1\,\mu m$ and an MRR with an FSR of 80.5 nm was realized [270]. As one might notice, sufficient coupling between the bus waveguide and the micro-ring waveguide is not easy because of the phase mismatch and the higher order modes excitation [266]. More recently, a low-loss add-drop optical filter with a record FSR as large as 93 nm was realized by using an MRR based on a *multi-mode* micro-ring waveguide (see Figure 29.16a and b), which was introduced to enable low-loss propagation even with a sub-micron bent radius (e.g., $R = 0.8\,\mu m$) [271]. More importantly, bent ADCs are introduced and designed optimally to minimize the excitation of the higher order modes in the *multi-mode* micro-ring waveguide. In this way, there is only one major resonance peak corresponding to the fundamental mode. In addition, this ultra-compact MRR has a SiO_2 upper-cladding other than air upper-cladding, which not only provides good reliability and mechanical stability but also makes it convenient for further development of advanced photonic integration. Alternatively, ultra-large FSR can be possibly achieved by using the Vernier effect in cascaded multiple MRRs with slightly different FSRs [272]. In this case, one usually needs to introduce the thermal-tuning technology to achieve precise alignment for resonance wavelengths. It is also possible to introduce some special filter inside the resonator to remove the resonance selectively and achieve only one major resonance [273,274]. For example, an FSR-free MRR-based optical filter was realized by using bent contra-DCs [273]. Unfortunately, the design and the fabrication for bent Bragg grating are not easy, and the device footprint is as large as ~1960 μm^2.

29.5.2.2 MRR-Based Optical Filters with Box-Like Responses

For many applications, it is desired to achieve optical filters with box-like spectral responses in order to tolerate some random resonance-wavelength shift due to random fabrication errors as well as random environmental variations. However, for single-ring micro-resonators, the Lorentzian-shaped spectral response does not have a flat top and sharp transitions. As demonstrated previously [275], box-like spectral responses have been realized by utilizing the structure of cascaded MRRs with optimized coupling ratios for all the couplers. In particular, the coupling ratio between the access waveguide and the side-micro-ring is often required to be sufficiently high [260], which can be obtained by using straight DCs as well as MMI couplers [260,261]. However, this prolongs the cavity length and prevents to have a maximal FSR. Furthermore, some EL is introduced due to the imperfect self-imaging in the MMI region or the mode-mismatching at the straight-bent junctions. Alternatively, bent DCs have been recognized as a preferred option for realizing cascaded MRRs with a box-like

TABLE 29.1

Comparison of Reported Ultra-Compact MRRs on Silicon

	Type	Radius (μm)	FSR (nm)	3dB-BW (nm)	Insertion Loss (dB)	Cladding
[266]	All-pass	1.5	62.5	0.17	NA	PMMA
[269]	Add-drop	2	47	0.24	NA	SiO_2
[267]	Add-drop	1.5	52	1.6	2.9	Air
[270]	Add-drop	1	80.5	3.3	0.85	Air
[271]	Add-drop	0.8	93	0.8	1.8	SiO_2

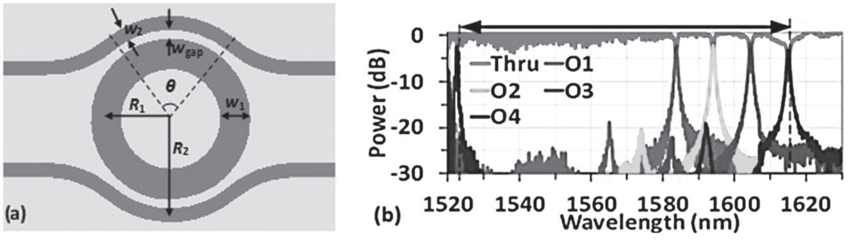


FIGURE 29.16 A sub-micron-ring resonator with bent ADCs: (a) schematic configuration; (b) measured spectral responses for an array with four channels. The record FSR is as large as 93 nm [271]. (Reprinted with permission from Ref. [271], [Optics Express].)

response since it was demonstrated for the first time in 2013 [263]. In this case, it is easy to achieve sufficient coupling ratio by choosing the length of the coupling region even when one chooses a relatively large gap-width for easy fabrication. Furthermore, the cavity length is minimized because no straight section is included in the micro-resonator. With this design, an ultra-compact MRR-based optical filter with a box-like spectral response and an FSR of >30 nm was realized for the first time [264] (Figure 29.17).

29.5.2.3 Polarization-Selective MRR Optical Filters

Regarding the applications with multiple wavelengths and dual polarizations (e.g. the realization of a crossing-free, compact, and scalable multi-wavelength coherent receiver), it is desired to have a kind of polarization-selective optical filters, which can select the wavelength channels of TE or TM polarization only [276]. As it is well known, both the coupling ratios of bent DCs and the bending loss in the micro-ring waveguide are strongly polarization sensitive. As a result, it is possible to a novel polarization-selective optical filter based on MRRs, as proposed in [276]. According to the transmission at the drop port of MRRs, the key to achieve polarization-selective MRRs is effectively manipulating the polarization-dependence of the bending loss and the coupling ratio for the MRRs, so that the critical coupling condition of a micro-resonator is satisfied for one polarization only and is broken for the other polarization automatically. In particular, the resonance of the undesired polarization can be depressed by minimizing the coupling ratio or enhancing the bending loss. Meanwhile, for the desired polarization mode, one can choose the coupling ratio and the bending loss optimally for achieving a low loss and a high ER. Such an MRR simultaneously serves as a high-ER polarization-handling device and an optical filter with low loss as well as low crosstalk.

In order to achieve strongly polarization-selective coupling, bent DCs are used for the MRRs and are designed according to the phase-matching condition [61]. For the TM-type MRR, which can drop the wavelength channel of TM polarization only, a regular micro-ring is used while the bent DC is designed to minimize the cross-coupling for TE polarization by introducing significant phase-mismatching, as shown in Figure 29.18a. For the TE-type MRR, which can drop the wavelength channel of TE polarization only, sharp waveguide bends with high bending losses for TM polarization are introduced, and the bent DCs are designed to minimize the coupling ratio for TM polarization, as shown in Figure 29.18b. As an example, Figure 29.19a and b show the measurement results for a TM-type MRR when the TE- and TM-polarization mode is launched, respectively. It can be seen that the realized TM-type MRR works as a regular filter with a low loss of <1 dB and a high extinction ratio of 20~25 dB for TM polarization. Meanwhile, this TM-type MRR allows TE-polarization mode to go through and finally output from the through port with a very low loss. The transmission at the drop port is <−50 dB for the TE-polarization mode over the wavelength range from 1530 to 1600 nm. As it can be seen, this realized TM-type MRR is highly polarization-selective and the polarization ER

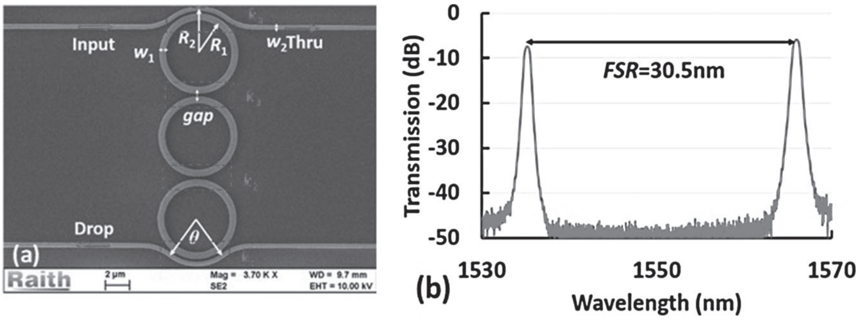


FIGURE 29.17 Cascaded MRRs with bent ADCs: (a) SEM picture; (b) measured spectral responses at the drop port [264]. (Copyright in 2018, Journal of Optics.)

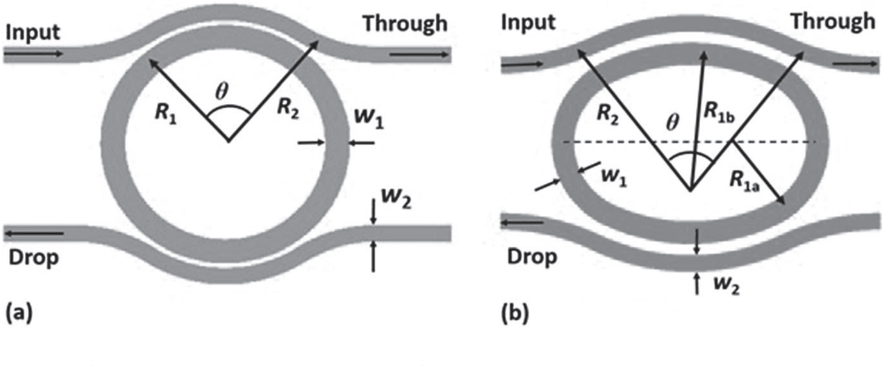


FIGURE 29.18 Configuration of polarization-selective MRR optical filters: (a) TM-type MRRs; (b) TE-type MRR. The TE-type MRR drops the wavelength channel of TE polarization only while the TM-type MRR drops the wavelength channel of TM polarization only [276]. (Reprinted with permission from Ref. [276], [Optics Express].)

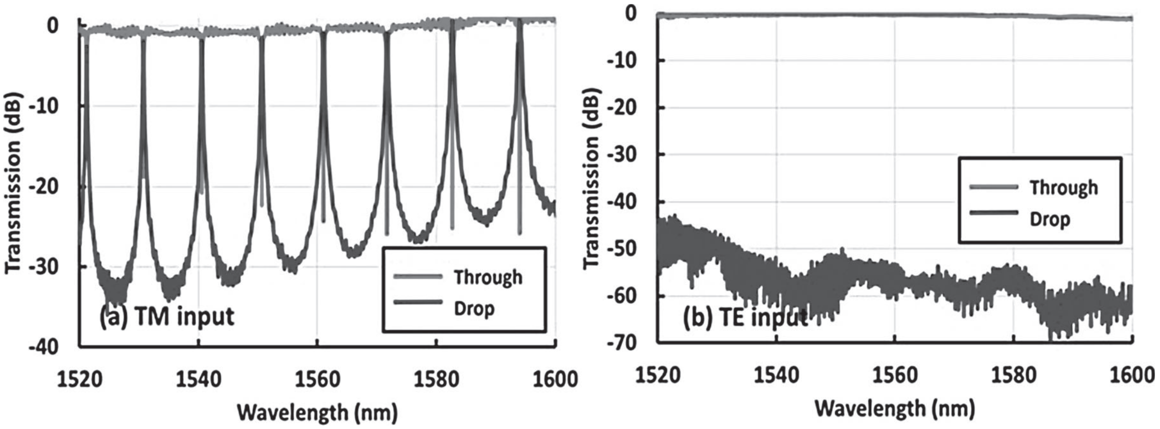


FIGURE 29.19 Measurement results for the TM-type MRR when TM-polarization mode (a) and TE-polarization mode (b) are launched, respectively [276]. (Reprinted with permission from Ref. [276], [Optics Express].)

at the drop port is >50 dB, which is useful for polarization-handling in WDM systems. This concept can be extended further to realize a polarization-selective arrayed MRRs with a box-like response, as demonstrated in [276].

29.5.3 Grating-Based Optical Filters

Waveguide gratings have attracted intensive attention because of the possibility of enabling the FSR-free optical filtering [277,278]. However, for traditional waveguide-grating filters, the problem is that only two ports are available (i.e. input/reflection, through). As a result, it is usually necessary to introduce circulators for separating the light reflected back to the input port, which makes the system very complicated. A promising approach is to develop add-drop optical filters with four ports (i.e. input, through, add, and drop) by using grating-assisted contra-directional couplers (contra-DCs), in which way the reflected light can be received at the drop port [279,280]. Recently a silicon contra-DC with an apodized SWG waveguide was demonstrated and realized with a high side-lobe suppression ratio (SLSR) of 27 dB [281]. Alternatively, another promising approach is using a multi-mode waveguide grating, as proposed in [282]. The grating was designed to enable the coupling from the forward fundamental mode (TE_0) to the backward first higher order mode (TE_1). The backward TE_1 mode is then coupled to the TE_0 mode at the drop port by using an adiabatic coupler. The multi-mode waveguide grating can be apodized longitudinally to improve the SLSR. For the fabricated device operating at 1.55 µm demonstrated in [282], the measured SLSR is about 7 dB and the bandwidth is about 3 nm for the transmission at the drop port. In [283], the SLSR was improved to 18.5 dB with the help of optimal apodization. This design was further improved by introducing grating tapers between the input/output waveguide and the grating section in order to reduce the EL, as shown in Figure 29.20a [284]. Some add-drop optical filters for other wavelength-bands (including 2 µm [284] and 1.31 µm [285]) have also been demonstrated. As demonstrated, the spectral response at the drop port has a low EL (~1.0 dB), a large 0.5 dB-bandwidth (6~26 nm), and a high SLSR (>20 dB). This type of optical filter was used further to achieve multi-channel WDM devices by cascading three MWG-based filters [285], as shown in Figure 29.20b. With this design, the on-chip triplexer has an EL of <0.5 dB and inter-channel crosstalk of <–45 dB (see Figure 29.20c–e).

29.6 Reconfigurable Silicon Photonic Devices and Circuits

The development of reconfigurable photonic integrated devices and circuits is important for making optical networks intelligent so that the bandwidth/channels can be utilized optimally and flexibly [286–301]. For silicon photonics, the most popular approaches for realizing optical switchability/tunability include the use of the electro-optic effect [296] or thermo-optic effect [297]. When using the electro-optic effect, one can achieve nanosecond-scale temporal responses. However, the accompanied free-carrier absorption is a barrier for the realization of low-loss large-scale reconfigurable silicon photonic devices and circuits (e.g. optical switch fabric). In contrast, when utilizing the thermo-optic effect, the response is relatively slow (i.e. ~µs) while the EL is negligible. Furthermore, silicon has a high thermo-optic coefficient (~1.8×10^{-4}/K @ $\lambda = 1.55\,\mu m$) as well as a high heat conductivity (~149 W/m·K) [297], and the heating volume is very small. Thus, it is possible to achieve high thermally tuning/switching efficiency and develop energy-efficient reconfigurable silicon photonic-integrated devices and circuits, which is the focus in the section.

29.6.1 Optical Switches

As a basic block to enable all-optical signal routing and switching in reconfigurable photonic networks/systems, a 2 × 2 optical switch has been developed successfully and there are two typical structures. One is based on MZIs and the other one is based on MRRs [298,299]. For MZI optical switches, it is often desired to achieve low losses as well as high ERs in a broadband covering all the wavelength channels in WDM systems [300]. In contrast, for MRR optical switches, which work wavelength-selectively, it is desired to have low losses and a high ERs within a narrow wavelength band.

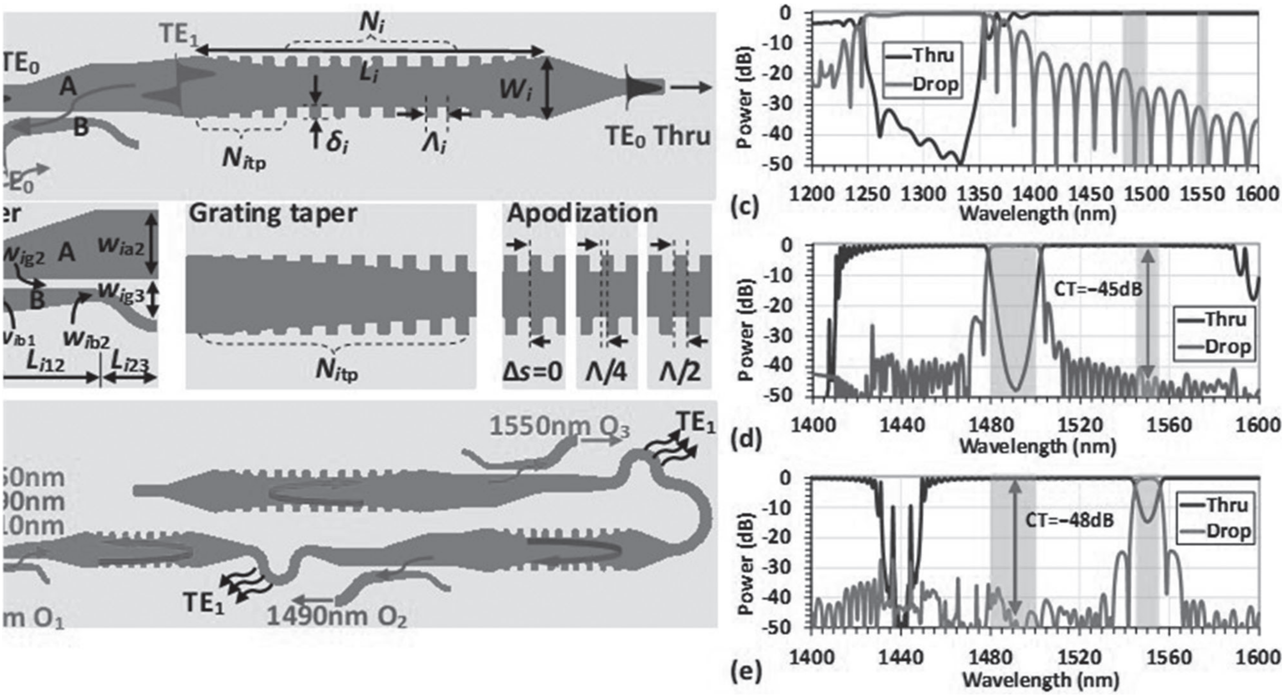


FIGURE 29.20 (a) The MWG-based optical filter (insets: the mode (de)multiplexer, the grating taper, Bragg grating); (b) optical triplexer; calculated transmissions of the designed triplexers for the channel of (c) 1310 nm, (d) 1490 nm, and (e) 1550 nm [285]. (Reprinted with permission from Ref. [285], [Optics Letter].)

In order to achieve broadband MZI optical switches, one should introduce 2 × 2 power splitters with a uniform coupling ratio of ~50%:50% over a broadband. As it is well known, for regular 2 × 2 straight DCs, the power splitting ratio is strongly wavelength-dependent [301]. Previously, 2 × 2 MMI couplers have been used widely as a better option regarding the footprint compactness and the relatively large fabrication tolerance. However, one should notice that a 2 × 2 MMI coupler has some intrinsic EL even in theory due to the imperfect self-imaging in an MMI section [302]. In [303], a 2 × 2 power splitter based on adiabatic DCs was demonstrated to enable the power-splitting ratio of ~50%:50% in a bandwidth of >100 nm. The problem is that the device is usually as long as ~100 μm to be adiabatic. Another improved design was demonstrated by using an MZI consisting of two DCs and asymmetric phase-arms [304,305]. In this case, it is needed to control the coupling ratios of the DCs and the phase difference of the MZI's arms very carefully. Furthermore, the optical switch is very long and complex, as demonstrated in [301]. As an improved design, a straight MZI consisting of two asymmetric phase-control sections was proposed and a 32-μm-long 2 × 2 3-dB power splitter was realized for TE polarization [306]. More recently, bent DCs were proposed to obtain broad-band power splitters (see Figure 29.21f) [307,308]. For the bent DCs designed optimally, the coupling ratio is 45%:55%~50%:50% in the wavelength band from 1518 to 1666 nm, which enables ultra-broadband MZI optical switches with an ER of >20 dB, as demonstrated in Figure 29.22a and b [309]. Here, these two bent DCs are arranged point-symmetrically to make the two MZI arms balanced. The measurement results show that the EL is very low (<0.5 dB) and the ER is >20 dB over a broadband of ~140 nm for both on/off states. This bandwidth is much larger than that for traditional MZI optical switches with straight DCs (whose bandwidth is ~35 nm [301]). It is also possible to realize polarization-insensitive MZI optical switches by introducing polarization-insensitive 2 × 2 3-dB power splitters based on cascaded bent DCs [310] as well as MMI couplers [311], as proposed recently.

For MRR optical switches, the wavelength-selectivity and wavelength-tunability naturally make it attractive for low power-consumption reconfigurable WDM systems with many wavelength channels [312]. As shown in Figure 29.23a, the resonance wavelength shifts due to the thermo-optic effect when the electrical power is applied. In this way, one can drop any wavelength channel flexibly as desired. Such tunable/switchable silicon MRRs can be cascaded conveniently to realize spatially non-blocking optical *N* × *N* router for future integrated photonic inter-connection networks [313,314]. On the other hand, when using a single micro-ring with Lorentzian spectral responses, the resonance wavelength should be aligned to the corresponding channel wavelength accurately by controlling temperature very well. Therefore, it is helpful when using cascaded MRRs with box-like spectral responses. As demonstrated in [295], a high-performance wavelength-insensitive optical switch based on cascaded MRRs was realized by optically modifying the resonance wavelength of the central micro-ring. Thermal switching is also available by using a micro-heater [263], as shown in Figure 29.23b. It can be seen that a high ER of 35 dB is achieved for the wavelength of 1550 nm and the thermally switching efficiency is about 0.1 nm mW^{-1} [263]. One can improve the thermally switching efficiency by shrinking the micro-rings, or introducing some special designs (e.g. thermal-isolation grooves, or suspended waveguides [260], or transparent graphene nano-heaters [315–317]).

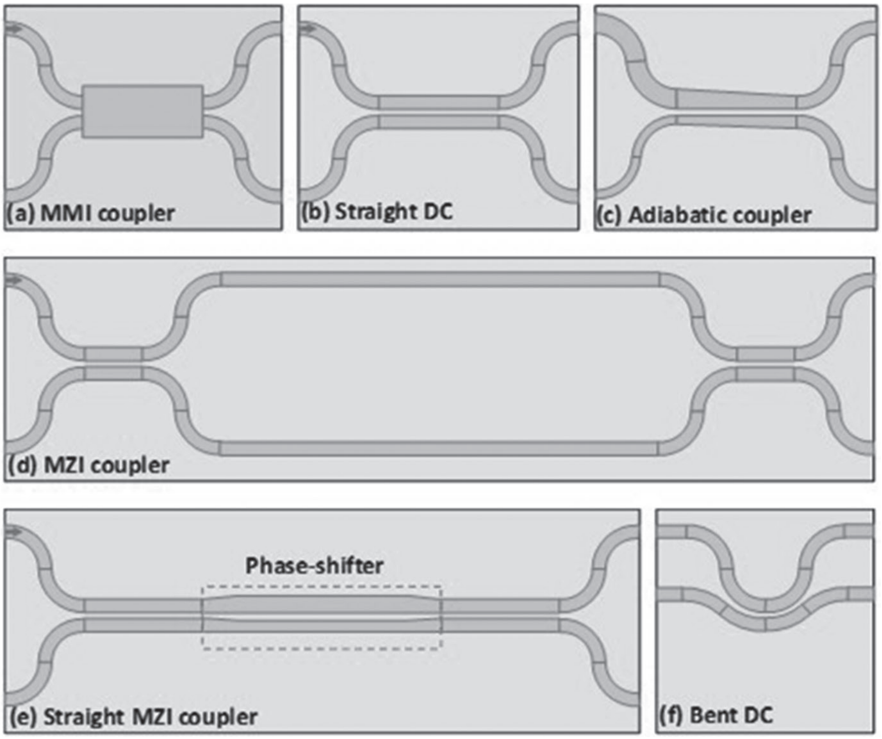

(a) MMI coupler
(b) Straight DC
(c) Adiabatic coupler
(d) MZI coupler
Phase-shifter
(e) Straight MZI coupler
(f) Bent DC

FIGURE 29.21 2 × 2 power splitters for MZI optical switches [19]. (Copyright in 2018, IEEE.) (a) MMI coupler [302]; (b) conventional straight DC [301]; (c) adiabatic DC [303]; (d) MZI coupler [311]; (e) straight MZI coupler [306]; (f) bent DC [309].

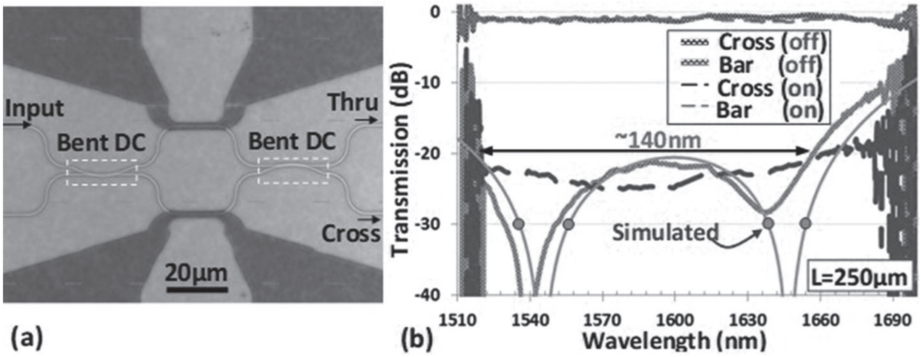


FIGURE 29.22 An MZI optical switch with bent DCs: (a) microscopy picture; (b) measurement results [309]. (Reprinted with permission from Ref. [309], [Optics Letter].)

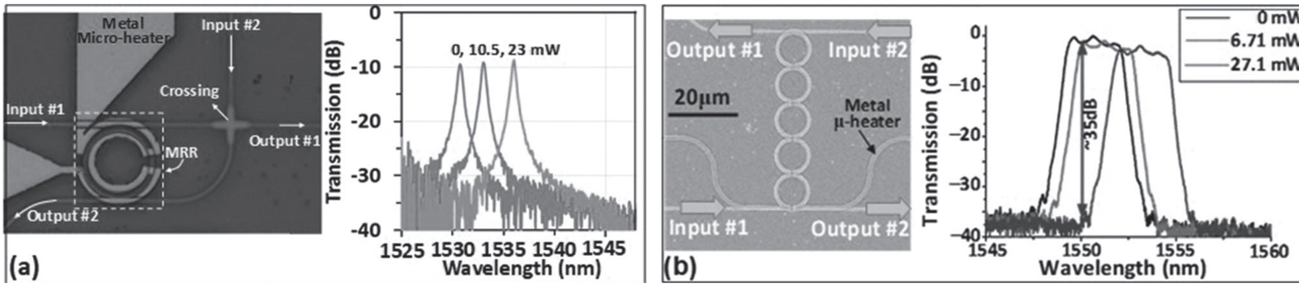


FIGURE 29.23 MRR-based wavelength-selective optical switches and the measured spectral responses as the electric power increases. (a) The switch with a single MRR [191]. (Reprinted with permission from Ref. [191], [Optics Letter].) (b) The switch with cascaded MRRs [263]. (Reprinted with permission from Ref. [263], [Optics Letter].)

29.6.2 ROADMs

ROADMs have been attracting much attention in the past decades as a key module for all-optical data-routing/switching at the nodes in multiplexed optical systems [318–326]. Typically, a ROADM consisting of a multiplexer, a demultiplexer and an array of optical switches. For example, the ROADM for WDM systems can be realized by using cascaded tunable MRRs with different resonance wavelengths [327,328], as shown in Figure 29.24a, where there are four-channel cascaded MRRs with bent DCs (here the micro-heaters are not shown). When tuning the resonance wavelength of the corresponding MRRs, one can add/drop any one of the wavelength channels flexibly. However, it might not be available for many wavelength channels due to the limited FSR [328]. Besides, the resonances for all MRRs are needed to be aligned to the channel wavelength very carefully, which makes the fabrication and the management not easy. As alternatives, EDGs and AWGs are usually preferred for the realization of ROADMs for WDM systems. In [329], a compact ROADM integrating four optic switches and two identical 4-channel EDGs was demonstrated for CWDM systems with channel spacing of 20 nm. In contrast, AWGs are often chosen for realizing DWDM-ROADMs. Recently a compact ROADM for 8 × 400 GHz WDM systems was realized by integrating two AWGs and eight thermo-optic MZI switches [330], as shown in Figure 29.24b. For these ROADMs with two AWGs/EDGs, some EL is introduced due to the wavelength misalignment between them. This issue can be alleviated by using flat-top AWGs/EDGs with a broadened 1-dB-bandwidth [331].

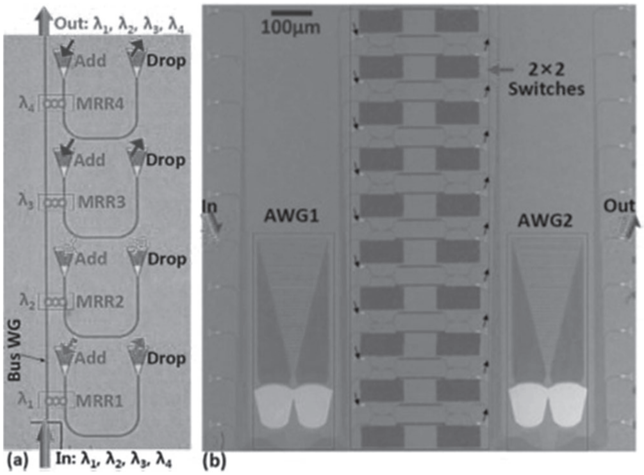


FIGURE 29.24 Silicon-based compact ROADM for WDM systems based on (a) MRRs [16]. (Reprinted with permission from Ref. [16], [Optics Express]) and (b) AWGs [330]. (Copyright in 2016, IEEE.)

Similarly, for MDM systems developed in recent years, it is also desired to have ROAMs to route or switch the data carried by different mode-channels, which can be realized by integrating mode (de)multiplexers and optical switches [332–334]. For example, a two-channel switchable data exchange for MDM systems was realized with a thermally switchable MZI consisting of multi-mode input/output waveguides [333]. In [332], a 2 × 2 multi-mode optical switch was realized by integrating two-channel ADC-based mode (de)multiplexers and an array of 2 × 2 single-mode optical switches. In [334], a high-speed (~2.5 ns) mode-selective electro-optic switch on silicon was demonstrated. In order to add or drop any mode-channels flexibly, a silicon-based ROADM chip was demonstrated recently, as shown in Figure 29.25 [190]. For this ROADM, a pair of four-channel mode (de)multiplexer (#1 and #2) and four thermo-optic MZI switches operating with the fundamental mode were integrated on a single chip. With this MDM-ROADM, any one of the mode-channels can be added or dropped individually by thermally turning on/off the corresponding MZI optical switch.

As it is well known, currently hybrid multiplexing systems have been proposed and thus it is also desired to develop the corresponding ROADMs. In [335], an on-chip multi-mode optical switch for hybrid WDM-MDM systems was demonstrated to manipulate two mode-channels and two wavelength channels by using two thermally tunable MRR-assisted ADCs in cascade. The data routing in hybrid WDM-MDM systems was also realized by combining grating-assisted ADCs and

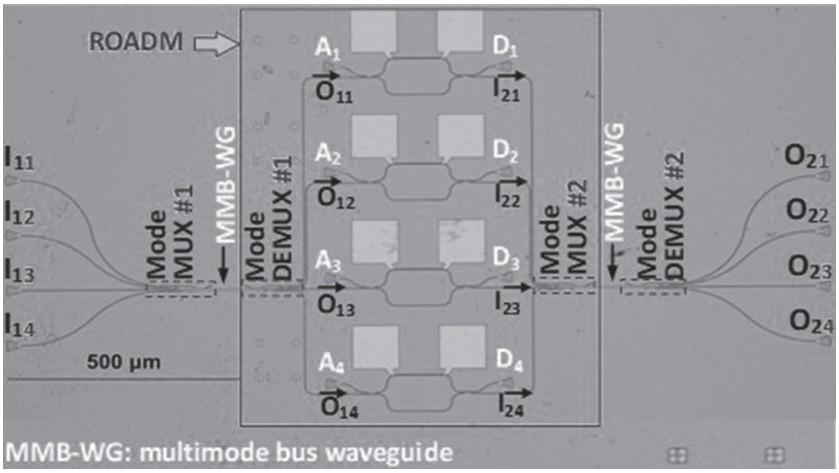


FIGURE 29.25 A four-channel ROADM for MDM systems [190]. (Reprinted with permission from Ref. [190], [Optics Letter].)

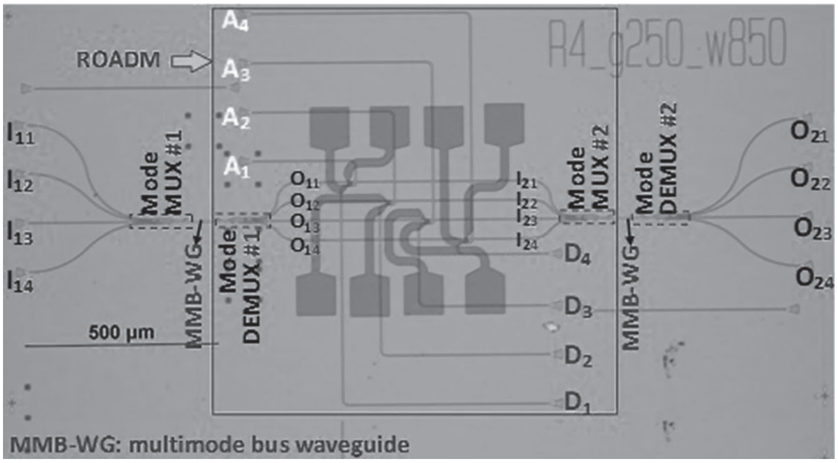


FIGURE 29.26 A ROADM for hybrid WDM-MDM systems [191]. (Reprinted with permission from Ref. [191], [Optics Letter].)

optical switches, as proposed in [332]. Figure 29.26a shows a silicon-based ROADM for hybrid WDM-MDM systems, realized by integrating a pair of mode (de)multiplexers and MRR-based wavelength-selective optical switches operating with the fundamental mode [191]. Here, another pair of mode (de)multiplexers was introduced at the input/output ends in order to characterize the ROADM conveniently. By using this ROADM, any wavelength channel of any mode can be added to (or dropped from) the multi-mode bus waveguide freely. More recently, a silicon-based ROADM for hybrid PDM-MDM systems was also demonstrated by using a 2 × 2 mode- and polarization-selective switch [192]. One should note that these ROADMs can be extended for manipulating more channels of wavelengths, modes as well as polarizations. For example, one can integrate the 10-channel hybrid PDM-MDM (de)multiplexers with MRR-based wavelength-selective optical switch arrays. In this way, one can add and drop the channels of any wavelength, any polarization, and any modes, which will be useful for the future multi-dimensional hybrid multiplexing systems.

29.7 Conclusion and Future Trends

In summary, this chapter has given a review of recent progresses on passive silicon photonic devices and circuits for optical communications. As discussed earlier, silicon nanophotonic waveguides have some special mode properties, e.g. ultra-high birefringence, high mode dispersion, and strong-mode hybridness. It has been shown that these special mode properties have been utilized very widely for the development of silicon photonic devices, which are very different from those traditional low-Δ optical waveguides.

First, structural asymmetry becomes extremely important for versatile silicon photonic devices. For example, ADCs (including bent DCs) have been recognized as a versatile element because the evanescent coupling can be controlled well by modifying the waveguide dimensions. As demonstrated, various excellent silicon photonic devices with asymmetric waveguide structures have been developed successfully for polarization/mode/wavelength/power manipulations, including on-chip polarization-handling devices (e.g. PBSs, PSRs), mode converters/(de)multiplexers, MRRs, etc. More importantly, these silicon photonic devices work very well with excellent performances, e.g. low loss, low crosstalk, ultra-broad bandwidth, and ultra-small footprints. In addition, the ADC structure can be extended for realizing dense waveguide array with non-uniform core widths, as demonstrated in [336–338]. Such dense waveguide arrays could be useful for multi-path on-chip data transmissions [338], sidelobeless on-chip beam-forming [339] as well as high-resolution on-chip micro-spectrometers [340], etc. It will be interesting to further explore the utilization of ADCs in silicon photonic devices in the future. More importantly, it is desired to develop large-scale PICs with these building blocks to satisfy the increasing demand of intelligent high-capacity data communications. For example, it is possible to achieve more than 1000 channels by combining multi-wavelengths, multi-modes as well as dual-polarizations. In order to satisfy this demand, on-chip hybrid (de)multiplexers and on-chip ROADMs will play more and more important roles in the near future.

Second, multi-mode silicon photonics has been very popular and higher order modes become key helpers for some applications since the manipulation of higher order modes in silicon nanophotonic waveguides becomes possible. Currently, multi-mode silicon photonic devices for MDM systems have been developed successfully with high performances (e.g. low loss, low crosstalk, compact footprints), including on-chip multi-channel mode (de)multiplexers, multi-mode waveguide bends, multi-mode waveguide crossings, reconfigurable multi-mode silicon PICs as well as multi-mode chip-fibre couplers. Some multi-mode optical switches have also been demonstrated in the past years [341,342]. Another important part of multi-mode silicon photonics is higher-order-mode-assisted silicon photonic devices, including on-chip polarization-handling devices with higher order modes, add-drop optical filters based on multi-mode Bragg gratings, etc. Some of these devices cannot be realized without the assistance of higher order modes. As it can be seen, the introduction of higher order modes provides more flexibility for the device design, which is beyond the single-mode condition. In addition, utilization of higher order modes in multi-mode silicon photonic waveguides also open the door for many applications. Currently, it has been recognized as a useful approach for lots of new applications of silicon photonics, e.g. quantum photonics [28–30], non-linear photonics [343], etc. Meanwhile, it is still desired to improve the performance of multi-mode silicon photonic devices to satisfy the demands for many applications. For example, more effort should be made to improve the multi-mode fibre-chip coupling so that the multi-mode silicon photonics can help the few-mode fibre optical communications more. The combination of three material platforms of SOI, α-Si, and Si_3N_4 might be helpful to create some hybrid waveguide structures with modified birefringence and mode dispersion for special mode-manipulations.

Third, further reduction of the power consumption and the propagation loss in silicon PICs is more and more important. In order to lower the power consumption, some novel waveguide structures and materials should be introduced. As an example, transparent nano-heaters based on, for example, graphene might be helpful because of the improved heating efficiency [315–317]. MEMS-based silicon photonic devices also provide a good option for achieving ultra-low power consumption [286]. It is also possible to develop some non-volatile silicon photonic devices with the help of phase-change materials [344,345]. Therefore, in order to develop powerful silicon PICs in the future, it is useful to introduce other functional materials "X" and versatile structures on silicon, e.g. low-loss SiN waveguides, highly nonlinear Lithium Niobate [346,347], and chalcogenide-glass waveguides [348]. Such a silicon-plus platform is very necessary to realize the monolithic integration of passive and active silicon photonic devices to enable large-scale silicon PICs. It also helps a lot for exploring more applications of silicon photonics. For example, it is becoming attractive to develop silicon photonic devices for the applications in those wavelength bands beyond the near-infrared band, e.g. optical sensing in the visible as well as mid-infrared ranges [349]. The development of quantum photonics on silicon has also attracted intensive attention in recent years [350] and will be more and more attractive in the future.

REFERENCES

1. M. J. R. Heck, J. F. Bauters, M. L. Davenport, et al. "Hybrid silicon photonic integrated circuit technology," *IEEE J. Sel. Top. Quantum Electron.*, vol. 19, no. 4, pp. 6100117–6100117, 2013.
2. M. Smit, J. van der Tol, and M. Hill. "Moore's law in photonics," *Laser Photon. Rev.*, vol. 6, no. 1, pp. 1–13, 2012.
3. M. Smit, X. Leijtens, H. Ambrosius, et al., "An introduction to InP-based generic integration technology," *Semiconductor Sc. Technol.*, vol. 29, no. 8, pp. 083001, 2014.
4. D. Thomson et al., "Roadmap on silicon photonics," *J. Opt.*, vol. 18, no. 7, pp. 073003, 2016.
5. R. Soref, "The past, present, and future of silicon photonics," *IEEE J. Sel. Top. Quant. Electron.*, vol. 12, no. 6, pp. 1678–1687, 2006.
6. B. Jalali and S. Fathpour, "Silicon photonics," *J. Lightwave Technol.*, vol. 24, no. 12, pp. 4600–4615, 2006.
7. C. Kopp et al., "Silicon photonic circuits: On-CMOS integration, fiber optical coupling, and packaging," *IEEE J. Sel. Top. Quant. Electron.*, vol. 17, no. 3, pp. 498–509, 2011.
8. L. Tsybeskov, D. Lockwood, and M. Ichikawa, "Silicon photonics: CMOS going optical," *Proc. IEEE*, vol. 97, no. 7, pp. 1161–1165, 2009.
9. M. Hochberg, T. Baehr-Jones, "Towards fabless silicon photonics," *Nat. Photon.*, vol. 4, no. 8, pp. 492–494, 2010.
10. E.-J. Lim et al., "Review of silicon photonics foundry efforts," *IEEE J. Sel. Top. Quant. Electron.*, vol. 20, no. 4, p. 8300112, 2014.
11. W. Bogaerts et al., "Silicon-on-insulator spectral filters fabricated with CMOS technology," *IEEE J. Sel. Top. Quant. Electron.*, vol. 16, no. 1, pp. 33–44, 2010.
12. M. Streshinsky et al., "The road to affordable, large-scale silicon photonics," *Opt. Photon. News*, vol. 24, no. 9, pp. 32–39, 2013.
13. P. Dumon et al., "Low-loss SOI photonic wires and ring resonators fabricated with deep UV lithography," *IEEE Photon. Technol. Lett.*, vol. 16, no. 5, pp. 1328–1330, 2004.
14. S. Selvaraja et al., "Fabrication of photonic wire and crystal circuits in silicon-on-insulator using 193nm optical lithography," *J. Lightwave Technol.*, vol. 27, no. 18, pp. 4076–4083, Sep. 2009.
15. D. Dai, Y. Shi, and S. He, "Characteristic analysis of nanosilicon rectangular waveguides for planar light-wave circuits of high integration," *Appl. Opt.*, vol. 45, no. 20, pp. 4941–4946, 2006.
16. D. Dai and J. E. Bowers, "Novel concept for ultracompact polarization splitter-rotator based on silicon nanowires," *Opt. Express*, vol. 19, pp. 10940–10949, 2011.
17. D. Dai and M. Zhang, "Mode hybridization and conversion in silicon-on-insulator nanowires with angled sidewalls," *Opt. Express*, vol. 23, pp. 32452–32464, 2015.
18. D. Dai, Y. Tang, and J. E. Bowers, "Mode conversion in tapered submicron silicon ridge optical waveguides," *Opt. Express*, vol. 20, no. 12, pp. 13425–13439, 2012.
19. D. Dai, "Advanced passive silicon photonic devices with asymmetric waveguide structures," *Proc. IEEE*, vol. 106, no. 12: 2117–2143, 2018.
20. D. Vermeulen, S. Selvaraja, W. A. D. De Cort, N. A. Yebo, E. Hallynck, K. De Vos, P. P. P. Debackere, P. Dumon, W. Bogaerts, G. Roelkens, D. Van Thourhout, and R. Baets, "Efficient tapering to the fundamental Quasi-TM mode in asymmetrical waveguides," ECIO*, 2010*, 2010.
21. T. Barwicz et al., "Polarization-transparent microphotonic devices in the strong confinement limit," *Nat. Photon.*, vol. 1, pp. 57–60, 2007.
22. W. Bogaerts, D. Taillaert, P. Dumon, D. Van Thourhout, and R. Baets, "A polarization-diversity wavelength duplexer circuit in silicon-on-insulator photonic wires," *Opt. Express*, vol. 15, pp. 1567–1578, 2007.
23. H. Fukuda, K. Yamada, T. Tsuchizawa, T. Watanabe, H. Shinojima, and S. Itabashi, "Silicon photonic circuit with polarization diversity," *Opt. Express*, vol. 16, pp. 4872–4880, 2008.
24. P. Dong, Y. K. Chen, G. H. Duan, and D. T. Neilson, "Silicon photonic devices and integrated circuits," *Nanophotonics*, vol. 3, no. 4–5, pp. 215–228, 2014.

25. T. Pfau et al., "Coherent digital polarization diversity receiver for real-time polarization-multiplexed QPSK transmission at 2.8 Gb/s," *IEEE Photon. Technol. Lett.* vol. 19, pp. 1988–1990, 2007.
26. P. Dong, X. Liu, S. Chandrasekhar, L. L. Buhl, R. Aroca, and Y. K. Chen, "Monolithic silicon photonic integrated circuits for compact 100(+)gb/s coherent optical receivers and transmitters," *IEEE J. Sel. Top. Quant. Electron.* vol. 20, no. 4, pp. 150–157, 2014.
27. P. Dong, "Silicon photonic integrated circuits for wavelength-division multiplexing applications," *IEEE J. Sel. Top. Quant. Electron.*, vol. 22, no. 6, pp. 370–378, 2016.
28. L. Feng et al., "On-chip coherent conversion of photonic quantum entanglement between different degrees of freedom," *Nat. Comm.* vol. 7, pp. 11985, 2016.
29. Mohanty A, Zhang M, Dutt A, et al., "Quantum interference between transverse spatial waveguide modes," *Nat Comm.* vol. 8, p. 14010, 2017.
30. L. Feng, M. Zhang, X. Xiong, Y. Chen, H. Wu, M. Li, G. P. Guo, G. Guo, D. Dai, and X. Ren, "On-chip transverse-mode entangled photon pair source," *NPG Quantum Inf.*, vol. 5, pp. 2, 2019.
31. D. Dai, L. Liu, S. Gao, D.-X. Xu, and S. He. "Polarization management for silicon photonic integrated circuits," *Laser Photon. Rev.* vol. 7, no. 3, pp. 303–328, 2013.
32. D. Dai, J. Bauters, and J. E Bowers, "Passive technologies for future large-scale photonic integrated circuits on silicon: polarization handling, light non-reciprocity, and loss reduction," *Light Sci. Appl.*, vol. 1, pp. 1–12, 2012.
33. J. M. Hong et al., "Design and fabrication of a significantly shortened multimode interference coupler for polarization splitter application," *IEEE Photon. Technol. Lett.*, vol. 15, no. 1, pp. 72–74, 2003.
34. Y. Jiao, D. Dai, Y. Shi, and S. He, "Shortened polarization beam splitters with two cascaded multimode interference sections," *IEEE Photon. Technol. Lett.*, vol. 21, no. 20, pp. 1538–1540, 2009.
35. Y. Huang, Z. Tu, H. Yi, Y. Li, X. Wang, and W. Hu, "High extinction ratio polarization beam splitter with multimode interference coupler on SOI," *Opt. Comm.*, vol. 307, pp. 46–49, 2013.
36. X. Guan, H. Wu, Y. Shi, and D. Dai, "Extremely small polarization beam splitter based on a multimode interference coupler with a silicon hybrid plasmonic waveguide," Opt. Lett., vol. 39, no. 2, pp. 259–262, 2014.
37. Y. Xu, J. Xiao, and X. Sun, "Compact polarization beam splitter for silicon-based slot waveguides using an asymmetrical multimode waveguide," *J. Lightwave Technol.*, vol. 32, no. 24, pp. 4282–4288, 2014.
38. A. Hosseini et al., "Ultracompact and fabrication-tolerant integrated polarization splitter," *Opt. Lett.*, vol. 36, no. 20, pp. 4047–4049, 2011.
39. Y. Ding, H. Ou, and C. Peucheret, "Wideband polarization splitter and rotator with large fabrication tolerance and simple fabrication process," *Opt. Lett.*, vol. 38, no. 8, pp. 1227–1229, 2013.
40. Y. Huang, Z. Tu, H. Yi, et al., "Polarization beam splitter based on cascaded step-size multimode interference coupler," *Opt. Eng.*, vol. 52, no. 7, pp. 077103–077103, 2013.
41. W. Yang et al., "A compact and wide-band polarization beam splitter based on wedge-shaped MMI coupler in silicon-on-insulator," *Optical Fiber Communication Conference*, Los Angeles, CA, USA, pp. W2A. 12, 2015.
42. M. Yin et al., "CMOS-compatible and fabrication-tolerant MMI-based polarization beam splitter," *Opt. Comm.*, vol. 335, pp. 48–52, 2015.
43. X. Sun, J. S. Aitchison, and M. Mojahedi, "Realization of an ultra-compact polarization beam splitter using asymmetric MMI based on silicon nitride/silicon-on-insulator platform," *Opt. Express*, vol. 25, no. 7, pp. 8296–8305, 2017.
44. B. K. Yang, S. Y. Shin, and D. M. Zhang, "Ultrashort polarization splitter using two-mode interference in silicon photonic wires," *IEEE Photon. Technol. Lett.*, vol. 21, no. 7, pp. 432–434, 2009.
45. L. B. Soldano, A. H. Vreede, M. K. Smit, B. H. Verbeek, E. G. Metaal, and F. H. Green, "Mach-Zehnder interferometer polarization splitter in InGaAsP-InP," *IEEE Photon. Technol. Lett.*, vol. 6, no. 3, pp. 402–405, 1994.
46. D. Dai, Z. Wang, J. Peters, and J. E. Bowers, "Compact polarization beam splitter using an asymmetrical Mach-Zehnder interferometer based on silicon-on-insulator waveguides," *IEEE Photon. Technol. Lett.*, vol. 24, no. 8, pp. 673–675, 2012.
47. T. K. Liang and H. K. Tsang, "Integrated PBS in high index contrast silicon-on-insulator waveguide," *IEEE Photon. Technol. Lett.*, vol. 17, no. 2, pp. 393–395, 2005.
48. L. M. Augustin, R. Hanfoug, Van der Tol, J. J. G. M., De Laat, W. J. M., and Smit, M. K., "A compact integrated polarization splitter/converter in InGaAsP–InP," *IEEE Photon. Technol. Lett.*, vol. 19, no. 17, pp. 1286–1288, 2007.
49. Y. Zhang et al., "High-extinction-ratio silicon polarization beam splitter with tolerance to waveguide width and coupling length variations," *Opt. Express*, vol. 24, no. 6, pp. 6586–6593, 2016.
50. Y. Shi, D. Dai, and S. He, "Proposal for an ultra-compact PBS based on a photonic crystal-assisted multimode interference coupler," *IEEE Photon. Technol. Lett.*, vol. 19, no. 11, pp. 825–827, 2007.
51. E. Cassan et al., "Polarization beam splitting using a birefringent graded photonic crystal," Opt. Lett., vol. 38, no. 4, pp. 459–461, 2013.
52. Y. Tang, D. Dai, and S. He, "Proposal for a grating waveguide serving as both a polarization splitter and an efficient coupler for silicon-on-insulator nanophotonic circuits," *IEEE Photon. Technol. Lett.*, vol. 21, no. 4, pp. 242–244, 2009.
53. X. Ao, L. Liu, L. Wosinski, and S. He, "Polarization beam splitter based on a two-dimensional photonic crystal of pillar type," *Appl. Phys. Lett.*, vol. 89, p. 171115, 2006.
54. M. R. Watts, H. A. Haus, and E. P. Ippen, "Integrated mode-evolution-based polarization splitter," *Opt. Lett.*, vol. 30, no. 9, pp. 967–969, 2005.
55. W. N. Ye, D. X. Xu, S. Janz, P. Waldron, P. Cheben, and N. G. Tarr, "Passive broadband silicon-on-insulator polarization splitter," *Opt. Lett.*, vol. 32, pp. 1492–1494, 2007.
56. Q. Z. Huang, Y. D. Yu, and J. Z. Yu, "Design and realization of a microracetrack resonator based polarization splitter in silicon-on-insulator," *J. Opt. A.*, vol. 11, p. 015506, 2009.

57. I. Kiyat, A. Aydinli, and N. A. Dagli, "Compact silicon-on-insulator polarization splitter," *IEEE Photon. Technol. Lett.*, vol. 17, no. 1, pp. 100–102, 2005.
58. L. M. Augustin, J. J. G. M. van der Tol, R. Hanfoug, W. J. M. de Laat, Van, M. J. E. de Moosdijk, M. J. E. Van de Moosdijk, P. W. L. van Dijk, Y. Oei, and M. Smit, "A single etch-step fabrication-tolerant polarization splitter," *J. Lightwave Technol*. vol. 25, no. 3, pp. 740–746, 2007.
59. Xiao, X. Liu, and X. Sun, "Design of a compact polarization splitter in horizontal multiple-slotted waveguide structures," *Jap. J. Appl. Phys.*, vol. 47, no. 5, pp. 3748–3754, 2008.
60. H. Fukuda et al., "Ultrasmall polarization splitter based on silicon wire waveguides," *Opt. Express*, vol. 14, pp. 12401–12408, 2006.
61. D. Dai and J. E. Bowers, "Novel ultra-short and ultra-broadband polarization beam splitter based on a bent directional coupler," *Opt. Express*, vol. 19, pp. 18614–18620, 2011.
62. M. A. Komatsu, K. Saitoh, and M. Koshiba, "Design of miniaturized silicon wire and slot waveguide polarization splitter based on a resonant tunneling," *Opt. Express*, vol. 17, pp. 19225–19234, 2009.
63. S. Lin, J. Hu, and K.B. Crozier, "Ultracompact, broadband slot waveguide polarization splitter," *Appl. Phys. Lett.*, vol. 98, pp. 151101, 2011.
64. D. Dai, Z. Wang, and J. E. Bowers, "Ultrashort broadband polarization beam splitter based on an asymmetrical directional coupler," *Opt. Lett.*, vol. 36, no. 13, pp. 2590–2592, 2011.
65. J. Feng, R. Akimoto, and H. Zeng, "Asymmetric silicon slot-waveguide-assisted polarizing beam splitter," *IEEE Photon. Tech. Lett.*, vol. 28, no. 12, pp. 1294–1297, 2016.
66. J. Wang, D. Liang, Y. Tang, D. Dai, and J. E. Bowers, "Realization of an ultra-short silicon polarization beam splitter with an asymmetrical bent directional coupler," *Opt. Lett.*, vol. 38, no. 1, pp. 4–6, 2013.
67. D. Chen et al., "Broadband, fabrication-tolerant polarization beam splitters based on a tapered directional coupler," *IEEE Photon. Technol. Lett.*, vol. 28, no. 19, pp. 2074–2077, 2016.
68. Z. Lu, Y. Wang, F. Zhang, N. A. F. Jaeger, and L. Chrostowski, "Wideband silicon photonic polarization beam splitter based on point-symmetric cascaded broadband couplers," *Opt. Express*, vol. 23, no. 23, 29413–29422, 2015.
69. D. W. Kim, M. H. Lee, Y. Kim, and K. H. Kim, "Planar-type polarization beam splitter based on a bridged silicon waveguide coupler," *Opt. Express*, vol. 23, no. 2, pp. 998–1005, 2015.
70. C.-W. Hsu, T.-K. Chang, J.-Y. Chen, and Y.-C. Cheng, "8.13 μm in length and CMOS compatible polarization beam splitter based on an asymmetrical directional coupler," *Appl. Opt.*, vol. 55, pp. 3313–3318, 2016.
71. F. Lou, D. Dai, and L. Wosinski, "Ultracompact polarization beam splitter based on a dielectric-hybrid plasmonic-dielectric coupler," *Opt. Lett.*, vol. 37, no. 16, pp. 3372–3374, 2012.
72. J. Chee, S. Zhu, and G. Q. Lo, "CMOS compatible polarization splitter using hybrid plasmonic waveguide," *Opt. Express*, vol. 20, no. 23, pp. 25345–25355, 2012.
73. X. Guan, H. Wu, Y. Shi, L. Wosinski, and D. Dai, "Ultracompact and broadband polarization beam splitter utilizing the evanescent coupling between a hybrid plasmonic waveguide and a silicon nanowire," *Opt. Lett.*, vol. 38, no. 16, pp. 3005–3008, 2013.
74. D. Dai, "Silicon polarization beam splitter based on an asymmetrical evanescent coupling system with three optical waveguides," *J. Lightwave Technol.*, vol. 30, no. 20, pp. 3281–3287, 2012.
75. S. Chen, H. Wu, and D. Dai, "High extinction-ratio compact polarisation beam splitter on silicon," *Electron. Lett.*, vol. 52, no. 12, pp. 1043–1045, 2016.
76. T. Y. Ang, J. R. Ong, E. Sahin, B. Pawlina, G. F. R. Chen, and D. T. H. Tan, "High-performance silicon-on-insulator polarization beam splitter using novel bend directional coupler," *The 13th International Conference on Group IV Photonics*, Shanghai, China, pp. 100–101, 2016.
77. Y. Zhang et al., "Ultra-compact broadband silicon polarization beam splitter based on a bridged bent directional coupler," *The 13th International Conference on Group IV Photonics*, Shanghai, China, pp. 116–117, 2016.
78. T. Y. L. Ang, J. R. Ong, S. T. Lim, B. Pawlina, E. Sahin, C. E. Png, H. S. Chu, G. F. R. Chen, and D. T. H. Tan, "Broadband silicon bridge waveguide polarization beam splitter," *Proc. SPIE* 10108, 101080A, 2017.
79. H. Wu and D. Dai, "High-performance polarizing beam splitters based on cascaded bent directional couplers," *IEEE Photon. Technol. Lett.*, vol. 29, no. 5, pp. 474–477, 2017.
80. S. Kim and M. Qi, "Copper nanorod array assisted silicon waveguide polarization beam splitter," *Opt. Express*, vol. 22, pp. 9508–9516, 2014.
81. B. Shen, P. Wang, R. Polson, and R. Menon, "An integrated-nanophotonics polarization beam splitter with 2.4×2.4 μm^2 footprint," *Nat. Photon.*, vol. 9, no. 6, pp. 378–382, 2015.
82. H. Wu, Y. Tan, and D. Dai, "Ultra-broadband high-performance polarizing beam splitter on silicon," *Opt. Express*, vol. 25, pp. 6069–6075, 2017.
83. Y. Xiong, J. G. Wangüemert-Pérez, D. X. Xu, J. H. Schmid, P. Cheben, and N. Y. Winnie, "Polarization splitter and rotator with subwavelength grating for enhanced fabrication tolerance," *Opt. Lett.*, vol. 39, no. 24, pp. 6931–6934, 2014.
84. C. Li and D. Dai, "Compact polarization beam splitter for silicon photonic integrated circuits with a 340nm-thick silicon core-layer," *Opt. Lett.*, vol. 42, no. 21, pp. 4243–4246, 2017.
85. V. R. Almeida, Q. Xu, C. A. Barrios, and M. Lipson, "Guiding and confining light in void nanostructure," *Opt. Lett.*, vol. 29, no. 11, pp. 1209–1211, 2004.
86. P. Cheben, D. X. Xu, S. Janz, and A. Densmore. "Subwavelength waveguide grating for mode conversion and light coupling in integrated optics," *Opt. Express*, vol. 14, pp. 4695–702, 2006.
87. Y. Xu and J. Xiao, "Compact and high extinction ratio polarization beam splitter using subwavelength grating couplers," *Opt. Lett.*, vol. 41, pp. 773–776, 2016.
88. L. Liu, Q. Deng, and Z. Zhou, "Manipulation of beat length and wavelength dependence of a polarization beam splitter using a subwavelength grating," *Opt. Lett.*, vol. 41, pp. 5126–5129, 2016.

89. D. Dai and S. He, “A silicon-based hybrid plasmonic waveguide with a metal cap for a nano-scale light confinement,” *Opt. Express*, vol. 17, no. 19, pp. 16646–16653, 2009.
90. H. Xu and Y. Shi, “On-chip silicon TE-pass polarizer based on asymmetrical directional couplers,” *IEEE Photon. Technol. Lett.*, vol. 1, pp. 861–864, 2017.
91. Q. Xu, L. Chen, M. G. Wood, P. Sun, and R. M. Reano, “Electrically tunable optical polarization rotation on a silicon chip using Berry’s phase,” *Nat. Comm.*, vol. 5, p. 5337, 2014.
92. K. Chin and S. T. Ho, “Design and modeling of waveguide-coupled single-mode microring resonators,” *J. Lightwave Technol.*, vol. 16, no. 8, pp. 1433–1446, Aug. 1998.
93. X. Guan et al., “Low-loss ultracompact transverse-magnetic-pass polarizer with a silicon subwavelength grating waveguide,” *Opt. Lett.*, vol. 39, no. 15, pp. 4514–4517, 2014.
94. W. Zhou, Z. Cheng, B. Zhu, X. Sun, and H. K. Tsang, “Hyperuniform disordered network polarizers” *IEEE J. Sel. Top. Quant. Electron.* vol. 22, no. 6, p. 4901307, 2016.
95. X. Sun, M. Z. Alam, S. J. Wagner, J. S. Aitchison, and M. Mojahedi, “Experimental demonstration of a hybrid plasmonic transverse electric pass polarizer for a silicon-on-insulator platform,” *Opt. Lett.*, vol. 37, no. 23, pp. 4814–4816, 2012.
96. J. R. Ong, T. Y. L. Ang, E. Sahin, B. Pawlina, G. F. R. Chen, D. T. H. Tan, S. T. Lim, and C. E. Png, “Broadband silicon polarization beam splitter with a high extinction ratio using a triple-bent-waveguide directional coupler,” *Opt. Lett.* 42, 4450–4453, 2017.
97. Y. Shani, R. Alferness, T. Koch, U. Koren, M. Oron, B. I. Miller, and M. G. Young, “Polarization rotation in asymmetric periodic loaded rib waveguides,” *Appl. Phys. Lett.*, vol. 59, no. 11, pp. 1278–1280, 1991.
98. Y. Zhang, Y. He, X. Jiang, B. Liu, C. Qiu, Y. Su, and R. A. Soref, “Ultra-compact and highly efficient silicon polarization splitter and rotator,” *APL Photon.*, vol. 1, no. 9, p. 091304, 2016.
99. van del Tol, J. J. G. M., F. Hakimzadeh, J. W. Pedersen, D. Li, and H. van Brug, “A new short and low-loss passive polarization converter on InP,” *IEEE Photon. Technol. Lett.*, vol. 7, no. 1, pp. 32–34, 1995.
100. Y. Sun, Y. Xiong, and N. Y. Winnie, “Compact SOI polarization rotator using asymmetric periodic loaded waveguides,” *IEEE Photon. J.*, vol. 8, no. 1, pp. 1–8, 2016.
101. S. Obayya, B. Rahman, K. Grattan, and H. El-Mikati, “Improved design of a polarization converter based on semiconductor optical waveguide bends,” Appl. Opt., vol. 40, no. 30, pp. 5395–5401, 2001.
102. S. Obayya, B. Rahman, K. Grattan, and H. El-Mikati, “Beam propagation modeling of polarization rotation in deeply etched semiconductor bent waveguides,” *IEEE Photon. Technol. Lett.*, vol. 13, pp. 681–683, 2001.
103. B. Holmes and D. Hutchings, “Realization of Novel Low-Loss Monolithically Integrated Passive Waveguide Mode Converters,” *IEEE Photon. Technol. Lett.*, vol. 12, no. 3, pp. 43–45, 2006.
104. Z. Huan, R. Scarmozzino, G. Nagy, J. Steel, and R. M. Osgood, “Realization of a compact and single-mode optical passive polarization converter,” *IEEE Photon. Technol. Lett.*, vol. 12, pp. 317–319, 2000.
105. H. Deng, D. Yevick, C. Brooks, and P. E. Jessop, “Fabrication tolerance of asymmetric silicon-on-insulator polarization rotators,” *J. Opt. Soc. Am. A*, vol. 23, no. 7, pp. 1741–1745, 2006.
106. H. Deng, D. Yevick, C. Brooks, and P. E. Jessop, “Design rules for slanted-angle polarization rotators,” *IEEE J. Lightwave Technol.*, vol. 23, no. 1, pp. 432–445, 2005.
107. C. Brooks et al., “Passive silicon-on-insulator polarization-rotating waveguides,” *Opt. Eng.*, vol. 45, no. 4, pp. 044603, 2006.
108. B. Daryl, M. Michele, and F. K. Thomas, “Compact polarization rotators for integrated polarization diversity in InP-based waveguides”, *Opt. Lett.*, vol. 32, no. 15, pp. 2176–2178, 2007.
109. Z. Wang and D. Dai, “Ultrasmall Si-nanowire-based polarization rotator,” *J. Opt. Soc. Am. B*, vol. 25, pp. 747–753, 2008.
110. Y. Wakabayashi, T. Hashimoto, J. Yamauchi, and H. Nakano, “Short waveguide polarization converter operating over a wide wavelength range,” *J. Lightwave Technol.*, vol. 31, pp. 1544–1550, 2013.
111. V. Velasco et al., “Ultracompact polarization converter with a dual subwavelength trench built in a silicon-on-insulator waveguide,” *Opt. Lett.* vol. 37, pp. 365–367, 2012.
112. D. Vermeulen, S. Selvaraja, and P. Verheyen, “Silicon-on-insulator polarization rotator based on a symmetry breaking silicon overlay,” *IEEE Photon. Technol. Lett.*, vol. 24, no. 6, pp. 482–484, 2012.
113. H. Fukuda, K. Yamada, T. Tsuchizawa, T. Watanabe, H. Shinojima, and S. I. Itabashi, “Polarization rotator based on silicon wire waveguides,” *Opt. Express*, vol. 16, pp. 2628–2635, 2008.
114. L. Gao, Y. Huo, J. S. Harris, and Z. Zhou, “Ultra-compact and low-loss polarization rotator based on asymmetric hybrid plasmonic waveguide,” *IEEE Photon. Technol. Lett.*, vol. 25, no. 21, pp. 2081–2084, 2013.
115. L. Liu, Y. Ding, K. Yvind, and J. M. Hvam, “Silicon-on-insulator polarization splitting and rotating device for polarization diversity circuits,” *Opt. Express*, vol. 19, pp. 12646–12651, 2011.
116. K. Tan, Y. Huang, G. Q. Lo, C. Lee, and C. Yu, “Compact highly-efficient polarization splitter and rotator based on 90° bends,” *Opt. Express*, vol. 24, no. 13, pp. 14506–14512, 2016.
117. J. Zhang, M. Yu, G. Lo, and D. L. Kwong, “Silicon waveguide based mode-evolution polarization rotator,” *IEEE J. Sel. Top. Quant. Electron.*, vol. 16, no. 1, pp. 53–60, 2010.
118. K. Goi, et al., “Low-loss partial rib polarization rotator consisting only of silicon core and silica cladding,” *Opt. Lett.*, vol. 40, no. 7, pp. 1410–1413, 2015.
119. Y. Xiong, D. X. Xu, J. H. Schmid, P. Cheben, S. Janz, and N. Y. Winnie, “Fabrication tolerant and broadband polarization splitter and rotator based on a taper-etched directional coupler,” *Opt. Express*, vol. 22, no. 14, pp. 17458–17465, 2014.
120. Y. Xu, and J. Xiao, “Design of a compact and integrated TM-rotated/TE-through polarization beam splitter for silicon-based slot waveguides,” *Appl. Opt.* vol. 55, no. 3, pp. 611–618, 2016.
121. X. Tu, M. Li, J. Xing, H. Fu, and D. Geng, “Compact PSR based on an asymmetric bi-level lateral taper in an adiabatic directional coupler,” *J. Lightwave Technol.*, vol. 34, pp. 985–991, 2016.

122. Y. Xiong, D. X. Xu, J. H. Schmid, P. Cheben, S. Janz, and N. Y. Winnie, “Robust silicon waveguide polarization rotator with an amorphous silicon overlayer,” *IEEE Photon. J.*, vol. 6, no. 2, pp. 1–8, 2014.
123. L. Chen, C. R. Doerr, and Y. K. Chen, “Compact polarization rotator on silicon for polarization-diversified circuits,” *Opt. Lett.*, vol. 36, no. 4, pp. 469–471, 2011.
124. J. Wang, B. Niu, Z. Sheng, A. Wu, W. Li, X. Wang, S. Zou, M. Qi, and F. Gan, “Novel ultra-broadband polarization splitter-rotator based on mode-evolution tapers and a mode-sorting asymmetric Y-junction,” *Opt. Express*, vol. 22, pp. 13565–13571, 2014.
125. W. D. Sacher, Y. Huang, L. Ding, T. Barwicz, J. C. Mikkelsen, B. J. F. Taylor, G.-Q. Lo, and J. K. S. Poon, “Polarization rotator-splitters and controllers in a Si_3N_4-on-SOI integrated photonics platform,” *Opt. Express*, vol. 22, pp. 11167–11174, 2014.
126. L. Socci, V. Sorianello, and M. Romagnoli, “300 nm bandwidth adiabatic SOI polarization splitter-rotators exploiting continuous symmetry breaking,” *Opt. Express*, vol. 23, pp. 19261–19271, 2015.
127. D. Dai and H. Wu, “Realization of a compact polarization splitter-rotator on silicon,” *Opt. Lett.*, vol. 41, pp. 2346–2349, 2016.
128. J. Wang et al., “Proposal for fabrication-tolerant SOI polarization splitter-rotator based on cascaded MMI couplers and an assisted bi-level taper,” *Opt. Express*, vol. 22, no. 23, pp. 27869–27879, 2014.
129. H. Guan et al., “Ultracompact silicon-on-insulator polarization rotator for polarization-diversified circuits,” *Opt. Lett.*, vol. 39, pp. 4703–4706, 2014.
130. J. Guo and Y. Zhao, “Analysis of mode hybridness in tapered waveguides,” *IEEE Photon. Technol. Lett.*, vol. 27, no. 23, pp. 2441–2444, 2015.
131. H. Guan, Q. Fang, G. Q. Lo, and K. Bergman, “High-efficiency biwavelength polarization splitter-rotator on the SOI platform,” *IEEE Photon. Technol. Lett.*, vol. 27, no. 5, pp. 518–521, 2015.
132. J. Wang et al., “A silicon-on-insulator polarization diversity scheme in the mid-infrared,” *Opt. Express*, vol. 23, pp. 15029–15037, 2015.
133. H. Xu and Y. Shi, “Ultra-compact and highly efficient polarization rotator utilizing multi-mode waveguides,” *Opt. Lett.*, vol. 42, no. 4, pp. 771–774, 2017.
134. D. Liang and J. E. Bowers, “Recent progress in lasers on silicon,” *Nat. Photon.*, vol. 4, no. 8, p. 511, 2010.
135. C. Li, D. Liu, and D. Dai, “Multimode silicon photonics,” *Nanophotonics*, vol. 8, no. 2, pp. 227–247, 2019.
136. R. W. C. Vance and J. D. Love, “Asymmetric adiabatic multiprong for mode-multiplexed systems,” *Electron. Lett.*, vol. 29, p. 2134, 1993.
137. J. D. Love, R. W. C. Vance, and A. Joblin, “Asymmetric, adiabatic multipronged planar splitters,” *Opt. Quant. Electron.*, vol. 28, no. 4, pp. 353–369, 1996.
138. B. T. Lee and S. Y. Shin, “Mode-order converter in a multimode waveguide,” *Opt. Lett.*, vol. 28, pp. 1660–1662, 2003.
139. Y. Kawaguchi and K. Tsutsumi, “Mode multiplexing and demultiplexing devices using multimode interference couplers,” *Electron. Lett.*, vol. 38, no. 25, pp. 1701–1702, 2002.
140. T. Uematsu, Y. Ishizaka, Y. Kawaguchi, et al., “Design of a compact two-mode multi/demultiplexer consisting of multimode interference waveguides and a wavelength-insensitive phase shifter for mode-division multiplexing transmission,” *J Lightwave Technol.*, vol. 30, pp. 2421–2426, 2012.
141. S. Bagheri and W. M. J. Green, “Silicon-on-insulator modeselective add-drop unit for on-chip mode-division multiplexing,” *The 6th IEEE International Conference on Group IV Photonics*, San Francisco, CA, USA, pp. 166–168, 2009.
142. J. Xing, Z. Li, X. Xiao, J. Yu, and Y. Yu, “Two-mode multiplexer and demultiplexer based on adiabatic couplers,” *Opt. Lett.*, vol. 38, pp. 3468–3470, 2013.
143. J. Driscoll, R. Grote, B. Souhan, J. Dadap, M. Lu, and R. Osgood, “Asymmetric Y junctions in silicon waveguides for on-chip mode-division multiplexing,” *Opt. Lett.*, vol. 38, pp. 1854–1856, 2013.
144. T. Pan and S. Tseng, “Short and robust silicon mode (de) multiplexers using shortcuts to adiabaticity,” *Opt. Express*, vol. 23, no. 8, pp. 10405–10412, 2015.
145. D. Dai, “Silicon mode-(de)multiplexer for a hybrid multiplexing system to achieve ultrahigh capacity photonic networks- on-chip with a single-wavelength-carrier light,” *Asia Communications and Photonics Conference,* Guangzhou, China, pp. ATh3B.3, 2012.
146. D. Dai, J. Wang, and Y. Shi, “Silicon mode, (de) multiplexer enabling high capacity photonic networks-on-chip with a single-wavelength- carrier light,” *Opt. Lett.*, vol. 38, pp. 1422–1424, 2013.
147. H. Qiu, H. Yu, T. Hu, G. Jiang, H. Shao, P. Yu, J. Yang, and X. Jiang, “Silicon mode multi/demultiplexer based on multimode grating-assisted couplers,” *Opt. Express*, vol. 21, pp. 17904–17911, 2013.
148. L. Luo, N. Ophir, C. Chen, L. H. Gabrielli, C. Poitras, K. Bergmen, and M. Lipson, “WDM-compatible mode-division multiplexing on a silicon chip,” *Nat. Comm.*, vol. 5, p. 3069, 2014.
149. Y. Ding, J. Xu, F. Da Ros, B. Huang, H. Ou, and C. Peucheret, “On-chip two-mode division multiplexing using tapered directional coupler-based mode multiplexer and demultiplexer,” *Opt. Express*, vol. 21, pp. 10376–10382, 2013.
150. C. Li and D. Dai, “Low-loss and low-crosstalk multichannel mode (de)multiplexer with ultrathin silicon waveguides,” *Opt. Lett.*, vol. 42, no. 12, pp. 2370–2373, 2017.
151. N. Hanzawa, K. Saitoh, T. Sakamoto, T. Matsui, S. Tomita, and M. Koshiba, “Asymmetric parallel waveguide with mode conversion for mode and wavelength division multiplexing transmission,” *Optical Fiber Communication Conference*, Los Angeles, CA, USA, pp. OTu1I.4, 2012.
152. Y. Li, C. Li, C. Li, et al., “Compact two-mode (de) multiplexer based on symmetric Y-junction and multimode interference waveguides,” *Opt. Express*, vol. 22, no. 5, pp. 5781–5786, 2014.
153. L. Han, S. Liang, H. Zhu, et al., “Two-mode de/multiplexer based on multimode interference couplers with a tilted joint as phase shifter,” *Opt. Lett.*, vol. 40, no. 4, pp. 518–521, 2015.

154. D. Gonzalez-Andrade, J. G. Wanguemert-Perez, A. Velasco, et al., "Ultra-broadband mode converter and multiplexer based on sub-wavelength structures," *IEEE Photon. J.*, vol. 99, pp. 1–1, 2018.
155. M. Greenberg and M. Orenstein, "Multimode add-drop multiplexing by adiabatic linearly tapered coupling," *Opt. Express*, vol. 13, pp. 9381, 2005.
156. N. Riesen and J. D. Love, "Design of mode-sorting asymmetric Y-junctions," *Appl. Opt.*, vol. 51, p. 2778, 2012.
157. W. Chen, P. Wang, and J. Yang, "Mode multi/demultiplexer based on cascaded asymmetric Y-junctions," *Opt. Express*, vol. 21, p. 25113, 2013.
158. W. Chen, P. Wang, J. Yang, et al., "Silicon three-mode (de) multiplexer based on cascaded asymmetric Y junctions," Opt. Lett., vol. 41, no. 12, pp. 2851–2854, 2016.
159. Z. Zhang, Y. Yu, and S. Fu, "Broadband on-chip mode-division multiplexer based on adiabatic couplers and symmetric Y-junction," *IEEE Photon. J*, vol. 9, no. 2, pp. 1–6, 2017.
160. W. Chang, L. Lu, X. Ren, et al., "Ultra-compact mode (de) multiplexer based on subwavelength asymmetric Y-junction," *Opt. Express*, vol. 26, no. 7, pp. 8162–8170, 2018.
161. J. Wang, Y. Xuan, M. Qi, et al., "Broadband and fabrication-tolerant on-chip scalable mode-division multiplexing based on mode-evolution counter-tapered couplers," *Opt. Lett.*, vol. 40, no. 9, pp. 1956–1959, 2015.
162. C. L. Sun, Y. Yu, G. Y. Chen, et al., "Silicon mode multiplexer processing dual-path mode-division multiplexing signals," *Opt. Lett.*, vol. 41, p. 5511, 2016.
163. L. F. Frellsen, Y. Ding, O. Sigmund, et al., "Topology optimized mode multiplexing in silicon-on-insulator photonic wire waveguides," *Opt. Express*, vol. 24, p. 16866, 2016.
164. S. Martínez-Garaot, S. Y. Tseng, and J. G. Muga, "Compact and high conversion efficiency mode-sorting asymmetric Y junction using shortcuts to adiabaticity," *Opt. Lett.*, vol. 39, no. 8, pp. 2306–2309, 2014.
165. S. Y. Tseng, R. D. Wen, Y. F. Chiu, et al., "Short and robust directional couplers designed by shortcuts to adiabaticity," *Opt. Express*, vol. 22, no. 16, pp. 18849–18859, 2014.
166. C. S. Y. Tseng, "Optimization of adiabaticity in coupled-waveguide devices using shortcuts to adiabaticity," *Opt. Lett.*, vol. 40, no. 21, pp. 4831–4834, 2015.
167. D. Guo and T. Chu, "Silicon mode (de) multiplexers with parameters optimized using shortcuts to adiabaticity," *Opt. Express*, vol. 25, no. 8, pp. 9160–9170, 2017.
168. D. Dai, C. Li, S. Wang, et al., "10-channel mode (de) multiplexer with dual polarizations," *Laser Photon. Rev.*, vol. 12, no. 1, pp. 1700109, 2018.
169. D. Dai, J. Wang, and Y. Shi, "Silicon mode (de) multiplexer enabling high capacity photonic networks-on-chip with a single-wavelength-carrier light," *Opt. Lett.*, vol. 38, no. 9, pp. 1422–1424, 2013.
170. D. Dai and S. P. Wang, "Asymmetric directional couplers based on silicon nanophotonic waveguides and applications," *Front. Optoelectron.*, vol. 9, pp. 450–465, 2016.
171. D. Dai and J. Wang, "Multi-channel silicon mode, (de) multiplexer based on asymmetrical directional couplers for on-chip optical interconnects," *IEEE Photon. Soc. News*, vol. 28, pp. 8–14, 2014.
172. M. Yin, Q. Deng, Y. Li, X. Wang, and H. Li, "Compact and broadband mode multiplexer and demultiplexer based on asymmetric plasmonic–dielectric coupling," *Appl. Opt.*, vol. 53, pp. 6175–6180, 2014.
173. W. Jiang, F. Cheng, J. Xu, and H. Wan, "Compact and low-crosstalk mode (de)multiplexer using a triple plasmonic-dielectric waveguide-based directional coupler," *J. Opt. Soc. Am. B*, vol. 35, pp. 2532–2540, 2018.
174. L. H. Gabrielli, D. Liu, S. G. Johnson, et al., "On-chip transformation optics for multimode waveguide bends," *Nat. Comm.*, vol. 3, p. 1217, 2012.
175. D. X. Dai, "Multimode optical waveguide enabling microbends with low inter-mode crosstalk for mode-multiplexed optical interconnects," *Opt. Express*, vol. 22, no. 22, pp. 27524–27534, 2014.
176. X. Wu, W. Zhou, D. Huang, et al., "Low Crosstalk Bent Multimode Waveguide for On-chip Mode-Division Multiplexing Interconnects," CLEO: QELS Fundamental Science. Optical Society of America, 2018, pp. JW2A. 66.
177. X. Jiang, H. Wu, and D. Dai, "Low-loss and low-crosstalk multimode waveguide bend on silicon," *Opt. Express*, vol. 26, no. 13, pp. 17680–17689, 2018.
178. C. L. Sun, Y. L. Yu, G. Y. Chen, C. T. Sima, S. N. Fu, and X. L. Zhan, "A novel sharply bent silicon multimode waveguide with ultrahigh mode extinction ratio," *Optical Fiber Communications Conference*, Anaheim, CA, USA, pp. W2A.12, 2016.
179. C. L. Sun, Y. Yu, G. Y. Chen, and X. L. Zhang, "Ultra-compact bent multimode silicon waveguide with ultralow inter-mode crosstalk," *Opt. Lett.*, vol. 42, no. 15, pp. 3004–3007, 2017.
180. M. Teng, A. Al Noman, Y. J. Lee, et al., "A 3-Micron-Radius Bend for SOI TE0/TE1 Multiplexing," CLEO: Applications and Technology. Optical Society of America, 2018, pp. JW2A. 13.
181. H. Xu, Y. Shi, "Ultra-sharp multi-mode waveguide bending assisted with metamaterial-based mode converters," *Laser Photon. Rev.*, vol. 12, no. 3, pp. 1700240, 2018.
182. W. Chang, L. Lu, D. Liu, et al., "Ultra-compact silicon multi-mode waveguide bend based on subwavelength asymmetric Y-junction," *Optical Fiber Communication Conference*, San Diego, CA, USA, pp. Tu3A. 1, 2018.
183. Y. Liu, W. Sun, H. Xie, et al., "Very sharp adiabatic bends based on an inverse design," *Opt. Lett.*, vol. 43, no. 11, pp. 2482–2485, 2018.
184. H. Wu, C. Li, L. Song, H. Tsang, J. Bowers, and D. Dai, "Ultra-sharp multimode waveguide bends with subwavelength gratings," *Laser Photon. Rev.*, vol. 13, no. 2, p. 1800119, 2019
185. H. Xu and Y. Shi, "Dual-mode waveguide crossing utilizing taper-assisted multimode-interference couplers," *Opt. Lett.*, vol. 41, no. 22, pp. 5381–5384, 2016.
186. C. Sun, Y. Yu, and X. Zhang, "Ultra-compact waveguide crossing for a mode-division multiplexing optical network," *Opt. Lett.*, vol. 42, no. 23, pp. 4913–4916, 2017.
187. W. Chang, L. Lu, X. Ren, et al., "An Ultracompact Multimode Waveguide Crossing Based on Subwavelength Asymmetric Y-Junction," *IEEE Photon. J.*, vol. 10, no. 4, pp. 1–8, 2018.

188. W. Chang, L. Lu, X. Ren, et al., “Ultracompact dual-mode waveguide crossing based on subwavelength multimode-interference couplers,” *Photon. Res.*, vol. 6, no. 7, pp. 660–665, 2018.
189. H. Xu and Y. Shi, “Metamaterial-based Maxwell’s fisheye lens for multimode waveguide crossing,” *Laser Photon. Rev.*, vol. 12, no. 10, p. 1800094.
190. S. P. Wang, H. Wu, H. K. Tsang, and D. Dai, “Monolithically integrated reconfigurable add-drop multiplexer for mode-division-multiplexing systems,” *Opt. Lett.*, vol. 41, pp. 5298, 2016.
191. S. P. Wang, X. L. Feng, S. M. Gao, et al., “On-chip reconfigurable optical add-drop multiplexer for hybrid wavelength/mode-division-multiplexing systems,” *Opt. Lett.*, vol. 42, pp. 2802–2805, 2017.
192. Y. Zhang, Y. He, Q. Zhu, C. Qiu, and Y. Su, “On-chip silicon photonic 2 × 2 mode- and polarization-selective switch with low inter-modal crosstalk,” *Photon. Res.*, vol. 5, pp. 521–526, 2017.
193. L. Yang, T. Zhou, H. Jia, et al., “General architectures for on-chip optical space and mode switching,” Optica, vol. 5, no. 2, pp. 180–187, 2018.
194. B. Wohlfeil, G. Rademacher, C. Stamatiadis, et al., “A Two-dimensional fiber grating coupler on SOI for mode division multiplexing,” *IEEE Photon. Technol. Lett.*, vol. 28, no. 11, pp. 1241–1244, 2016.
195. Y. Lai, Y. Yu, S. Fu, et al., “Efficient spot size converter for higher-order mode fiber-chip coupling,” *Opt. Lett.*, vol. 42, no. 18, pp. 3702–3705, 2017.
196. Y. Lai, Y. Yu, S. Fu, et al., “Compact double-part grating coupler for higher-order mode coupling,” *Opt. Lett.*, vol. 43, no. 13, pp. 3172–3175, 2018.
197. D. Dai and M. Mao, “Mode converter based on an inverse taper for multimode silicon nanophotonic integrated circuits,” *Opt. Express*, vol. 23, no. 22, pp. 28376–28388, 2015.
198. Y. Wu and K. S. Chiang, “Ultra-broadband mode multiplexers based on three-dimensional asymmetric waveguide branches,” *Opt. Lett.*, vol. 42, no. 3, p. 407, 2017.
199. K. Chen, J. Y. Wu, X. Li, et al., “Broadband five-mode (de) multiplexer with horizontal tapered directional couplers,” *Asia Communications and Photonics Conference,* Guangzhou, China, pp. Su3K. 7, 2017.
200. Q. Huang, Q. Liu, and J. Xia, “Traveling wave-like Fabry–Perot resonator-based add-drop filters,” *Opt. Lett.*, vol. 42, no. 24, pp. 5158–5161, 2017.
201. Q. Huang,, K. Jie, Q. Liu, et al., “Ultra-compact, broadband tunable optical bandstop filters based on a multimode one-dimensional photonic crystal waveguide,” *Opt. Express*, vol. 24, no. 18, pp. 20542–20553, 2016.
202. X. Guan and L. H. Frandsen, “All-silicon thermal independent Mach-Zehnder interferometer with multimode waveguides,” *IEEE International Conference on Group IV Photonics*, Shanghai, China, pp. 8–9, 2016.
203. D. Dai, J. Wang, and S. He, “Silicon multimode photonic integrated devices for on-chip mode-division-multiplexed optical interconnects,” *Prog. Electromagn. Res.*, vol. 143, pp. 773–819, 2013.
204. J. Wang, S. He, and D. Dai, “On-chip silicon 8-channel hybrid (de)multiplexer enabling simultaneous mode- and polarization-division-multiplexing,” *Laser Photon. Rev.*, vol. 8, pp. 18–22, 2014.
205. Y. Yang, Y. Li, Y. Huang, et al., “Silicon nitride three-mode division multiplexing and wavelength-division multiplexing using asymmetrical directional couplers and microring resonators,” *Opt. Express*, vol. 22, p. 22172, 2014.
206. J. Wang, S. Chen, and D. Dai, “Silicon hybrid demultiplexer with 64 channels for wavelength/mode-division multiplexed on-chip optical interconnects,” *Opt. Lett.*, vol. 39, no. 24, pp. 6993, 2014.
207. J. Wang, P. Chen, S. Chen, et al., “Improved 8-channel silicon mode demultiplexer with grating polarizers,” *Opt. Express*, vol. 22, pp. 12799–12807, 2014.
208. L. S. Han, S. Liang, J. J. Xu, et al., “Simultaneous wavelength- and mode-division (de)multiplexing for high-capacity on-chip data transmission link,” *IEEE Photon. J*, vol. 8, pp. 1–10, 2016.
209. K. Ji and H. M. Chen, “A hybrid multiplexer for wavelength/mode-division based on photonic crystals,” *Proc. SPIE*, vol. 244, p. 102440B, 2017.
210. D. Dai, J. Wang, S. Chen, et al., “Monolithically integrated 64-channel silicon hybrid demultiplexer enabling simultaneous wavelength- and mode-division-multiplexing,” *Laser Photonics Rev*, vol. 9, pp. 339–344, 2015.
211. S. Wang, H. Wu, M. Zhang, and D. Dai, “A 32-channel hybrid wavelength-/mode-division (de)multiplexer on silicon,” *IEEE Photon. Technol. Lett.*, vol. 30, pp. 1194–1197, 2018.
212. Y. Tan, H. Wu, S. Wang, C. Li, and D. Dai, “Silicon-based hybrid demultiplexer for wavelength- and mode-division multiplexing,” *Opt. Lett.*, vol. 43, no. 9, p. 1962, 2018.
213. C. Li, H. Wu, Y. Tan, et al., “Silicon-based on-chip hybrid (de) multiplexers,” *Sci. China Inf. Sci.*, vol. 61, no. 8, p. 080407, 2018.
214. Y. Tan, H. Wu, and D. Dai, “Silicon-based hybrid (de)multiplexer for wavelength-/polarization-division- multiplexing,” *J. Lightwave Technol.*, vol. 36, no. 11, pp. 2051–2058, 2018.
215. E. Erol and H. S. Sözüer, “High transmission through a 90° bend in a polarization-independent single-mode photonic crystal waveguide,” *Opt. Express*, vol. 23, no. 25, pp. 32690–32695, 2015.
216. B. Shen, R. Polson, and R. Menon, “Metamaterial-waveguide bends with effective bend radius< $\lambda_0/2$,” *Opt. Lett.* vol. 40, no. 24, pp. 5750–5753, 2015.
217. T. Fujisawa, S. Makino, T. Sato, et al., “Low-loss, compact, and fabrication-tolerant Si-wire 90° waveguide bend using clothoid and normal curves for large scale PICs,” *Opt. Express*, vol. 25, no. 8, pp. 9150–9159, 2017.
218. W. Bogaerts and S. K. Selvaraja, “Compact single-mode silicon hybrid rib/strip waveguide with adiabatic bends,” *IEEE Photon. J.*, vol. 3, no. 3, pp. 422–432, 2011.
219. Y. A. Vlasov and S. J. McNab, “Losses in single-mode silicon-on-insulator strip waveguides and bends,” *Opt. Express*, vol. 12, no. 8, pp. 1622–1631, 2004.
220. M. Cherchi, S. Ylinen, M. Harjanne, et al., “Dramatic size reduction of waveguide bends on a micron-scale silicon photonic platform,” *Opt. Express*, vol. 21, no. 15, pp. 17814–17823, 2013.
221. D. Dai, Y. Shi, and S. He, “Comparative study of the integration density for passive linear planar light-wave circuits based on three different kinds of nanophotonic waveguide,” *Appl. Opt.*, vol. 46, no. 7, pp. 1126–1131, 2007.

222. K. Yamada, T. Shoji, T. Tsuchizawa, T. Watanabe, T. Takahashi, and S. Itabashi, "Silicon-wire-based ultrasmall lattice filters with wide free spectral ranges," *Opt. Lett*, vol. 28, pp. 1663–1664, 2003.
223. K. Jinguji and M. Oguma, "Optical half-band filters," *J Lightwave Technol.*, vol. 18, pp. 252–259, 2000.
224. M. Oguma, T. Kitoh, Y. Inoue, et al., "Compact and low-loss interleave filter employing lattice-form structure and silica-based waveguide," *J Lightwave Technol.*, vol. 22, pp. 895–902, 2004.
225. F. Horst, W. M. J. Green, S. Assefa, S. M. Shank, Y. A. Vlasov, and B. J. Offrein, "Cascaded Mach-Zehnder wavelength filters in silicon photonics for low loss and flat pass-band WDM (de-)multiplexing," *Opt. Express*, vol. 21, pp. 11652–11658, 2013.
226. K. Okamoto, "Progress and technical challenge for planar waveguide devices: silica and silicon waveguides," *Laser Photon Rev.*, vol. 6, no. 1, pp. 14–23, 2011.
227. H. Xu, Y. Shi, "Flat-top CWDM (de) multiplexer based on MZI with bent directional couplers," *IEEE Photon. Technol. Lett.*, vol. 30, no. 2, pp. 169–172, 2018.
228. C. Dragone, "An N×N optical multiplexer using a planar arrangement of two star couplers," *IEEE Photon. Technol. Lett.*, vol. 3, no. 9, pp. 812–814, 1991.
229. C. Doerr and K. Okamoto, "Advances in Silica Planar Lightwave Circuits," *J. Lightwave Technol.*, vol. 24, no. 12, pp. 4763–4789, 2006.
230. M. K. Smit and C. Dam, "Phasar-based WDM-devices: principles, design and applications," *IEEE J. Select. Top. Quant. Electron.*, vol. 2, no. 2, pp. 236–250, 1996.
231. S. N. Khan, D. Dai, L. Liu, L. Wosinski, and S. He, "Optimal design for a flat-top AWG demultiplexer by using a fast calculation method based on a Gaussian beam approximation," *Opt. Comm.*, vol. 262, no. 2, pp. 175–179, 2006.
232. P. Dumon, W. Bogaerts, D. V. Thourhout, et al. "Wavelength-selective components in SOI photonic wires fabricated with deep UV lithography," *Proceedings of Conference on Group IV Photonics*, Hong Kong, Sep. 2004, p. WB5.
233. T. Fukazawa, F. Ohno, and T. Baba, "Very compact arrayed-waveguidegrating demultiplexer using Si photonic wire waveguides," *Jpn. J. Appl. Phys.*, vol. 43, no. 5B, pp. 673–675, 2004.
234. K. Sasaki, F. Ohno, A. Motegi, and T. Baba, "Arrayed waveguide grating of 70×60μm^2 size based on Si photonic wire waveguides," *Electron. Lett.*, vol. 41, no. 14, pp. 801–802, 2005.
235. D. Dai, L. Liu, L. Wosinski, and S. He, "Design and fabrication of ultrasmall overlapped AWG demultiplexer based on alpha-SOI nanowire waveguides," *Electron. Lett.*, vol. 42, pp. 400–402, 2006.
236. W. Bogaerts, P. Dumon, D. Van Thourhout, et al., "Compact wavelength selective functions in silicon-on-insulator photonic wires," *J. Sel. Top. Quant. Electron.*, vol.12, no. 6, pp. 1394–1401, 2006.
237. P. Dumon, W. Bogaerts, D. Van Thourhout, et al., "Compact wavelength router based on a Silicon-on-insulator arrayed waveguide grating pigtailed to a fiber array," *Opt. Express*, vol. 14, pp. 664–669, 2006.
238. Q. Fang, T. Y. Liow, J. F. Song, et al., "WDM multi-channel silicon photonic receiver with 320 Gbps data transmission capability," *Opt. Express*, vol. 18, pp. 5106–5113, 2010.
239. J. Zou, Z. Le, J. Hu, and J. He, "Performance improvement for silicon-based arrayed waveguide grating router," *Opt. Express*, vol. 25, pp. 9963–9973, 2017.
240. T. Ye, Y. Fu, L. Qiao, and T. Chu, "Low-crosstalk Si arrayed waveguide grating with parabolic tapers," *Opt. Express*, vol. 22, pp. 31899–31906, 2014.
241. D. Dai and J. E Bowers, "Silicon-based on-chip multiplexing technologies and devices for Peta-bit optical interconnects," *Nanophotonics*, vol. 3, no. 4–5, pp. 283–311, 2014.
242. D. Dai, X. Fu, Y. Shi, and S. He, "Experimental demonstration of an ultra-compact Si-nanowire-based reflective arrayed-waveguide grating (de)multiplexer with photonic crystal reflectors," *Opt. Lett*, vol. 35, no. 15, pp. 2594–2596, 2010.
243. Z. Sheng, D. Dai, and S. He, "Improve channel uniformity of an Si-nanowire AWG demultiplexer by using dual-tapered auxiliary waveguides," *J. Lightwave Technol.*, vol. 25, no. 10, pp. 3001–3007, 2007.
244. G. Song, S. Wang, J. Zou, T. Lang, and J. J. He, "Silicon-based cyclic arrayed waveguide grating routers with improved loss uniformity," *Opt. Comm.*, vol. 427, pp. 628–634, 2018.
245. Y. Shi, X. Fu, and D. Dai. "Design and fabrication of a 200GHz Si-nanowire-based reflective arrayed-waveguide grating (de)multiplexer with optimized photonic crystal reflectors," *Appl Opt.*, vol. 49, no. 26, pp. 4859–4865, 2010.
246. X. Fu and D. Dai, "Ultra-small Si-nanowire-based 400 GHz-spacing 15×15 arrayed-waveguide grating router with microbends," *Electron. Lett.*, vol. 47, no. 4, pp. 266–268, 2011.
247. J. Brouckaert, W. Bogaerts, P. Dumon, D. Van Thourhout, and R. Baets, "Planar concave grating demultiplexer fabricated on a nanophotonic silicon-on-insulator platform," *J. Lightwave Technol.*, vol. 25, no. 5, pp. 1269–1275, 2007.
248. J. Brouckaert, W. Bogaerts, S. Selvaraja, P. Dumon, R. Baets, and D. Van. Thourhout, "Planar concave grating demultiplexer with high reflective bragg reflector facets," *IEEE Photon. Technol. Lett.*, vol. 20, no. 4, pp. 309–211, 2008.
249. F. Horst, "Silicon integrated waveguide devices for filtering and wavelength demultiplexing," *Optical Fiber Communication Conference*, Los Angeles, CA, USA, pp. OWJ3, 2011.
250. F. Horst, W. M. J. Green, B. J. Offrein, and Y. A. Vlasov, "Silicon-on-insulator echelle grating WDM demultiplexers with two stigmatic points," *IEEE Photon. Technol. Lett.*, vol. 21, no. 23, pp. 1743–1745, 2009.
251. B. B. Kyotoku, L. Chen, and M. Lipson, "Broad Band 1 nm channel spacing silicon-on-insulator wavelength division multiplexer," *Conference on Lasers and Electro-Optics/International Quantum Electronics Conference*, OSA Technical Digest (CD), Baltimore, Maryland, USA, pp. JWA41, 2009.
252. B. E. Little, S. T. Chu, H. A. Haus, and J. Foresi, "Microring resonator channel dropping filters," *J. Lightwave Technol.*, vol. 15, no. 6, pp. 998–1005, 1997.
253. PopovMiloš, T. Barwicz, M. R. Watts, et al., "Multistage high-order microring-resonator add-drop filters," *Opt. Lett.*, vol. 31, pp. 2571–2573, 2006.

254. H. Shen, M. H. Khan, L. Fan, et al., "Eight-channel reconfigurable microring filters with tunable frequency, extinction ratio and bandwidth," *Opt. Express*, vol. 18, pp. 18067–18076, 2010.
255. F. Xu and A. W. Poon, "Silicon cross-connect filters using microring resonator coupled multimode-interference-based waveguide crossings," *Opt. Express*, vol. 16, pp. 8649–8657, 2008.
256. M. K. Smit, "New focusing and dispersive planar component based on an optical phased array," *Electron. Lett.*, vol. 24, no. 7, pp. 385–386, 1988.
257. C. Dragone, "Efficient N × N star coupler based on Fourier optics," *Electron. Lett.*, vol. 24, no. 15, pp. 942–944, 1988.
258. H. Takahashi, S. Suzuki, K. Kato, and I. Nishi, "Arrayed-waveguide grating for wavelength division multi/demultiplexer with nanometre resolution," *Electron. Lett.*, vol. 26, no. 2, pp. 87–88, 1990.
259. L. Hansuek, C. Tong, Jiang L, et al., "Chemically etched ultrahigh-Q wedge-resonator on a silicon chip," *Nature Photon.*, vol. 6, pp. 369–373, 2012.
260. P. Dong, W. Qian, H. Liang, R. Shafiiha, D. Feng, G. Li, J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari, "Thermally tunable silicon racetrack resonators with ultralow tuning power," *Opt. Express*, vol. 18, pp. 20298–20304, 2010.
261. F. Xia, M. Rooks, L. Sekaric, and Y. Vlasov, "Ultra-compact high order ring resonator filters using submicron silicon photonic wires for on-chip optical interconnects," *Opt. Express*, vol. 15, 11934–11941, 2007.
262. J. Wang, and D. Dai, "Highly sensitive Si nanowire-based optical sensor using a Mach-Zehnder interferometer coupled microring," *Opt. Lett.*, vol. 35, no. 24, pp. 4229–4231, 2010.
263. P. Chen, S. Chen, X. Guan, Y. Shi, and D. Dai, "High-order microring resonators with bent couplers for a box-like filter response," *Opt. Lett.*, vol. 39, pp. 6304–6307, 2014.
264. Y. Tan, D. Dai, "Silicon microring resonators," *J. Opt.*, vol. 20, no. 5, pp. 054004, 2018.
265. J. Niehusmann, A. Vörckel, P. H. Bolivar, T. Wahlbrink, and W. Henschel, "Ultrahigh-quality-factor silicon-on-insulator microring resonator," *Opt. Lett.*, vol. 29, pp. 2861–2863, 2006.
266. Q. Xu, D. Fattal, and R. G. Beausoleil, "Silicon microring resonators with 1.5-μm radius," *Opt. Express*, vol. 16, pp. 4309–4315, 2008.
267. M. Prabhu, A. Tsay, Z. Han, and V. Van, "Ultracompact SOI microring add–drop filter with wide bandwidth and wide FSR," *IEEE Photon. Technol. Lett.*, vol. 21, no. 10, pp. 651–653, 2009.
268. Q. Xu, B. Schmidt, J. Shakya, and M. Lipson, "Cascaded silicon micro-ring modulators for WDM optical interconnection," *Opt. Express*, vol. 14, pp. 9431–9435, 2006.
269. M. S. Nawrocka, T. Liu, X. Wang, and R. R. Panepucci, "Tunable silicon microring resonator with wide free spectral range," *Appl. Phys. Lett.*, vol. 89, no. 7, pp. 071110, 2006.
270. M. Prabhu, A. Tsay, Z. Han, and V. Van, "Extreme miniaturization of silicon add–drop microring filters for VLSI photonics applications," *IEEE Photon. J.*, vol. 2, no. 3, pp. 436–444, 2010.
271. D. Liu, C. Zhang, D. Liang, and D. Dai, "Submicron-resonator-based add-drop optical filter with an ultra-large free spectral range," *Opt. Express*, vol. 27, pp. 416–422, 2019.
272. T. Claes, W. Bogaerts, and P. Bienstman, "Experimental characterization of a silicon photonic biosensor consisting of two cascaded ring resonators based on the Vernier-effect and introduction of a curve fitting method for an improved detection limit," *Opt. Express*, vol. 18, no. 22, pp. 22747–22761, 2010.
273. N. Eid, R. Boeck, H. Jayatilleka, L. Chrostowski, W. Shi, and N. A. F. Jaeger, "FSR-free silicon-on-insulator microring resonator based filter with bent contra-directional couplers," *Opt. Express*, vol. 24, no. 25, pp. 29009–29021, 2016.
274. D. Urbonas, A. Balčytis, M. Gabalis, K. Vaškevičius, G. Naujokaitė, S. Juodkazis, and R. Petruškevičius, "Ultrawide free spectral range, enhanced sensitivity, and removed mode splitting SOI optical ring resonator with dispersive metal nanodisks," *Opt. Lett.*, vol. 40, no. 13, pp. 2977–2980, 2015.
275. B. E. Little, S. T. Chu, J. V. Hryniewicz, and P. P. Absil, "Filter synthesis for periodically coupled microring resonators," *Opt. Lett.*, vol. 25, pp. 344–346, 2000.
276. Y. Tan, S. Chen, and D. Dai, "Polarization-selective microring resonators," *Opt. Express*, vol. 25, pp. 4106–4119, 2017.
277. H. Y. Qiu et al., "FSR-free add-drop filter based on silicon grating-assisted contra directional couplers," *Opt. Lett.*, vol. 38, no. 1, pp. 1–3, 2013.
278. W. F. Zhang and J. P. Yao, "A fully reconfigurable waveguide Bragg grating for programmable photonic signal processing," *Nat. Comm.*, vol. 9, p. 1396, 2018.
279. W. Shi et al., "Silicon photonic grating-assisted, contra-directional couplers," *Opt. Express*, vol. 21, no. 3, pp. 3633–3650, 2013.
280. B. Naghdi and L. R. Chen, "Silicon photonic contra directional couplers using subwavelength grating waveguides," *Opt. Express*, vol. 24, no. 20, pp. 23429–23438, Oct. 3 2016.
281. B. Y. Liu et al., "Silicon photonic bandpass filter based on apodized subwavelength grating with high suppression ratio and short coupling length," *Opt. Express*, vol. 25, no. 10, pp. 11359–11364, May 15 2017.
282. H. Y. Qiu et al., "Silicon add-drop filter based on multimode Bragg sidewall gratings and adiabatic couplers," *J. Lightwave Technol.*, vol. 35, no. 9, pp. 1705–1709, May 1 2017.
283. J. F. Jiang et al., "Silicon lateral-apodized add-drop filter for on-chip optical interconnection," *Appl. Opt.*, vol. 56, no. 30, pp. 8425–8429, 2017.
284. D. Liu, H. Wu, and D. Dai, "Silicon multimode waveguide grating filter at 2 μm," *IEEE J. Lightwave Technol.*, vol. 37, pp. 2217–2222, 2019.
285. D. Liu, M. Zhang, and D. Dai, "Low-loss and low-crosstalk silicon triplexer based on cascaded multimode waveguide gratings," *Opt. Lett.*, vol. 44, pp. 1304–1307, 2019.
286. T. J. Seok, N. Quack, S. Han, R. S. Muller, and M. C. Wu, "Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers," *Optica*, vol. 3, pp. 64–70, 2016.
287. M. Yang, W. M. J. Green, S. Assefa, et al., "Non-blocking 4x4 electro-optic silicon switch for on-chip photonic networks," *Opt. Express*, vol. 19, no. 1, pp. 47–54, 2011.

288. K. Tanizawa, K. Suzuki, M. Toyama, et al., “Ultra-compact 32× 32 strictly-non-blocking Si-wire optical switch with fan-out LGA interposer,” *Opt. Express*, vol. 23, no. 13, pp. 17599–17606, 2015.
289. C. M. Wilkes, X. Qiang, J. Wang, et al., “60 dB high-extinction auto-configured Mach–Zehnder interferometer,” *Opt. Lett.*, vol. 41, no. 22, pp. 5318–5321, 2016.
290. R. L. Espinola, M. C. Tsai, J. T. Yardley, et al., “Fast and low-power thermooptic switch on thin silicon-on-insulator,” *IEEE Photon. Technol. Lett.*, vol. 15, no. 10, pp. 1366–1368, 2003.
291. Y. Shoji, K. Kintaka, S. Suda, et al, “Low-crosstalk 2 × 2 thermo-optic switch with silicon wire waveguides,” *Opt. Express*, vol. 18, no. 9, pp. 9071–9075, 2010.
292. Q. Fang, J. F. Song, T. Y. Liow, et al., “Ultralow power silicon photonics thermo-optic switch with suspended phase arms,” *IEEE Photon. Technol. Lett.*, vol. 23, no. 8, pp. 525–527, 2011.
293. K. Suzuki, G. Cong, K. Tanizawa, et al., “Ultra-high-extinction-ratio 2 × 2 silicon optical switch with variable splitter,” *Opt. Express*, vol. 23, no. 7, pp. 9086–9092, 2015.
294. X. Wang, B. Howley, M. Y. Chen, et al., “4 × 4 nonblocking polymeric thermo-optic switch matrix using the total internal reflection effect,” *IEEE J. Sel. Top. Quant. Electron.*, vol. 12, no. 5, pp. 997–1000, 2006.
295. Y. Vlasov, W. M. J. Green, and F. Xia, “High-throughput silicon nanophotonic wavelength-insensitive switch for on-chip optical networks,” *Nat. Photon.*, vol. 2, no. 4, p. 242, 2008.
296. L. Zhou, L. Lu, Z. Li, S. Zhao, D. Li, and J. Chen, “Silicon large-scale optical switches using MZIs and dual-ring assisted MZIs,” *SPIE OPT*, San Diego, CA, USA, pp. 97520K-97520K, 2016.
297. J. Song, Q. Fang, S. H. Tao, et al., “Fast and low power Michelson interferometer thermo-optical switch on SOI,” *Opt. Express*, vol. 16, no. 20, pp. 15304–15311, 2008.
298. T. Chu, H. Yamada, S. Ishida, et al., “Compact 1 × N thermo-optic switches based on silicon photonic wire waveguides,” *Opt. Express*, vol. 13, no. 25, pp. 10109–10114, 2005.
299. J. Van Campenhout, W. M. J. Green, Y. A. Vlasov, “Design of a digital, ultra-broadband electro-optic switch for reconfigurable optical networks-on-chip,” *Opt. Express*, vol. 17, no. 26, pp. 23793–23808, 2009.
300. G. Calò, A. D‘Orazio, and V. Petruzzelli, “Broadband Mach-Zehnder switch for photonic networks on chip,” *J. Lightwave Technol.*, vol. 30, pp. 944–952, 2012.
301. J. Van Campenhout, W. M. Green, S. Assefa, et al., “Low-power, 2 × 2 silicon electrooptic switch with 110-nm bandwidth for broadband reconfigurable optical networks,” *Opt. Express*, vol. 17, pp. 24020–2402, 2009.
302. K. Suzuki, H. C. Nguyen, T. Tamanuki, F. Shinobu, Y. Saito, Y. Sakai, and T. Baba, “Slow-light-based variable symbol-rate silicon photonics DQPSK receiver,” *Opt. Express*, vol. 20, pp. 4796–4804, 2012.
303. H. Yun, W. Shi, Y. Wang, L. Chrostowski, and N. A. F. Jaeger, “2×2 adiabatic 3-db coupler on silicon-on-insulator rib waveguides,” *Proceedings of SPIE, Photonics North*, Ottawa, Canada, pp. 89150V-89150V-6, 2013.
304. K. Jinguji, N. Takato, A. Sugita, and M. Kawachi, “Mach-Zehnder interferometer type optical wave-guide coupler with wavelength-flattened coupling ratio,” *Electron. Lett.*, vol. 26, pp. 1326–1327, 1990.
305. A. Takagi, K. Jinguji, and M. Kawachi, “Wavelength characteristics of 2 × 2 optical channel-type directional-couplers with symmetrical or nonsymmetric coupling structures,” *J. Lightwave Technol.*, vol. 10, pp. 735–746, 1992.
306. Z. Lu, H. Yun, Y. Wang, Z. Chen, F. Zhang, N. A.F. Jaeger, and L. Chrostowski, “Broadband silicon photonic directional coupler using asymmetric-waveguide based phase control,” *Opt. Express*, vol. 23, pp. 3795–3806, 2015.
307. R. Doerr, M. Cappuzzo, E. Chen, A. Wong-Foy, L. Gomez, A. Griffin, and L. Buhl, “Bending of a planar lightwave circuit 2x2 coupler to desensitize it to wavelength, polarization, and fabrication changes,” *IEEE Photon. Technol. Lett.*, vol. 17, pp. 1211–1213, 2005.
308. H. Morino, T. Maruyama, and K. Iiyama, “Reduction of wavelength dependence of coupling characteristics using si optical waveguide curved directional coupler,” *J. Lightwave Technol.*, vol. 32, pp. 2188, 2014.
309. S. Chen, Y. Shi, S. He, and D. Dai, “Low-loss and broadband 2×2 silicon thermo-optic Mach-Zehnder switch with bent directional couplers,” *Opt. Lett.*, vol. 41, pp. 836–839, 2016.
310. X. Chen, W. Liu, Y. Zhang, and Y. Shi, “Polarization-insensitive broadband 2 × 2 3 dB power splitter based on silicon-bent directional couplers,” *Opt. Lett.*, vol. 42, pp. 3738–3740, 2017.
311. S. S. Agashe, K. T. Shiu, and S. R. Forrest, “Compact polarization-insensitive InGaAsP-InP 2 × 2 optical switch,” *IEEE Photon. Technol. Lett.*, vol. 17, no. 1, pp. 52–54, 2005.
312. K. Takahashi, Y. Kanamori, Y. Kokubun, and K. Hane, “A wavelength-selective add-drop switch using silicon microring resonator with a submicron-comb electrostatic actuator,” *Opt. Express*, vol. 16, pp. 14421–14428, 2008.
313. B. A. Small, B. G. Lee, K. Bergman, Q. Xu, and M. Lipson, “Multiple-wavelength integrated photonic networks based on microring resonator devices,” *J. Opt. Netw.*, vol. 6, pp. 112–120, 2007.
314. R. Ji, et al., “Microring-resonator-based four-port optical router for photonic networks-on-chip,” *Opt. Express*, vol. 19, pp. 18945–18955, 2011.
315. Q. L. Bao, and K. P. Loh, “Graphene photonics, plasmonics, and broadband optoelectronic devices,” *Acs Nano*, vol. 6, pp. 3677–3694, 2012.
316. Balandin, S. Ghosh, W. Z. Bao, I. Calizo, D. Teweldebrhan, F. Miao, and C. N. Lau, “Superior thermal conductivity of single-layer graphene,” *Nano Lett.*, vol. 8, pp. 902–907, 2008.
317. L. Yu, Y. Yin, Y. Shi, D. Dai, and S. He, “Thermally tuning silicon photonic micro-disk resonator with graphene transparent nano-heaters,” *Optica*, vol. 3, pp. 159–166, 2016.
318. L. Eldada, “Advances in ROADM technologies and subsystems,” *Proc. SPIE*, p. 5970B, 2005.
319. R. Min, R. Ji, Q. Chen, et al., “A universal method for constructing N-port nonblocking optical router for photonic networks-on-chip,” *J. Lightwave Technol.*, vol. 30, no. 23, pp. 3736–3741, 2012.
320. N. Sherwood-Droz, H. Wang, L. Chen, et al., “Optical 4 × 4 hitless silicon router for optical networks-on-chip (NoC),” *Opt. Express*, vol. 16, no. 20, pp. 15915–15922, 2008.
321. R. Ji, L. Yang, L. Zhang, et al., “Five-port optical router for photonic networks-on-chip,” *Opt. Express*, vol. 19, no. 21, pp. 20258–20268, 2011.

322. T. Hu, H. Qiu, P. Yu, et al., "Wavelength-selective 4 × 4 nonblocking silicon optical router for networks-on-chip," *Opt. Lett.*, vol. 36, no. 23, pp. 4710–4712, 2011.
323. H. Jia, T. Zhou, X. Fu, et al., "Four-port mode-selective silicon optical router for on-chip optical interconnect," *Opt. Express*, vol. 26, no. 8, pp. 9740–9748, 2018.
324. X. Tan, M. Yang, L. Zhang, et al., "A generic optical router design for photonic network-on-chips," *J. Lightwave Technol.*, vol. 30, no. 3, pp. 368–376, 2012.
325. T. Hu, H. Shao, L. Yang, et al., "Four-port silicon multi-wavelength optical router for photonic networks-on-chip," *IEEE Photon. Technol. Lett.*, vol. 25, no. 23, p. 2281, 2013.
326. R. Ji, J. Xu, and L. Yang, "Five-port optical router based on microring switches for photonic networks-on-chip," *IEEE Photon. Technol. Lett.*, vol. 25, no. 5, pp. 492–495, 2013.
327. M. Geng, L. Jia, L. Zhang, L. Yang, P. Chen, T. Wang, and Y. Liu, "Four-channel reconfigurable optical add-drop multiplexer based on photonic wire waveguide," *Opt. Express*, vol. 17, pp. 5502–5516, 2009.
328. E. J. Klein, et al., "Reconfigurable optical add-drop multiplexer using microring resonators," *IEEE Photon. Technol. Lett.*, vol. 17, no. 11, pp. 2358–2360, 2005.
329. J. Song, X. Luo, Q. Fang, L. Jia, X. Tu, T. Y. Liow, M. Yu, and G. Q. Lo, "Silicon-based low-power reconfigurable optical add-drop multiplexer, ROADM," *International Journal of Electronics and Communication Engineering*, vol. 6, no. 9, pp.945–948, 2012.
330. S. Chen, Y. Shi, S. He, and D. Dai, "Compact 8-channel thermally reconfigurable optical add/drop multiplexer on silicon," *IEEE Photon. Technol. Lett.*, vol. 28, no. 17, pp. 1874–1877, 2016.
331. S. Pathak, M. Vanslembrouck, P. Dumon, D. Thourhout, and W. Bogaerts, "Optimized silicon AWG with flattened spectral response using an MMI aperture," *J. Lightwave Technol.*, vol. 31, no. 1, pp. 87–93, 2013.
332. H. Jia, T. Zhou, L. Zhang, et al., "Optical switch compatible with wavelength division multiplexing and mode division multiplexing for photonic networks-on-chip," *Opt. Express*, vol. 25, no. 17, pp. 20698–20707, 2017.
333. C. Sun, Y. Yu, G. Chen, et al., "Integrated switchable mode exchange for reconfigurable mode-multiplexing optical networks," *Opt. Lett.*, vol. 41, no. 14, pp. 3257–3260, 2016.
334. Y. Xiong, R. B. Priti, and O Liboiron-Ladouceur, "High-speed two-mode switch for mode-division multiplexing optical networks," *Optica*, vol. 4, no. 9, pp. 1098–1102, 2017.
335. Grieco, G. Porter, and Y. Fainman, "Integrated space-division multiplexer for application to data center networks," *IEEE J. Sel. Top. Quant. Electron.*, vol. 22, no. 6, pp. 1–6, 2016.
336. S. Jahani, Kim S, Atkinson J, et al., "Controlling evanescent waves using silicon photonic all-dielectric metamaterials for dense integration," *Nature Comm.*, vol. 9, no. 1, pp. 1893, 2018.
337. W. Song, R. Gatdula, S. Abbaslou, et al., "High-density waveguide superlattices with low crosstalk," *Nature Comm.*, vol. 6, p. 7027, 2015.
338. L. Liu, "Densely packed waveguide array (DPWA) on a silicon chip for mode division multiplexing," *Opt. Express*, vol. 23, no. 9, pp. 12135–43, 2015.
339. D. N. Hutchison, J. Sun, J. K. Doylend, R. Kumar, J. Heck, W. Kim, C. T. Phare, A. Feshali, and H. Rong, "High-resolution aliasing-free optical beam steering," *Optica* vol. 3, pp. 887–890, 2016.
340. K. Ma, K. Chen, N. Zhu, et al., "High-resolution compact on-chip spectrometer based on an echelle grating with densely packed waveguide array," *IEEE Photon. J.*, vol. 11, no. 1, p. 4900107, 2019.
341. B. Stern, X. Zhu, C. P. Chen, et al., "On-chip mode-division multiplexing switch," *Optica*, vol. 2, no. 6, pp. 530–535, 2015.
342. Y. Zhang, Q. Zhu, Y. He, et al., "Silicon 1× 2 mode-and polarization-selective switch," *Optical Fiber Communication Conference,* Los Angeles, CA, USA, pp. W4E. 2, 2017.
343. E. A. Kittlaus, N. T. Otterstrom, and P. T. Rakich, "On-chip inter-modal Brillouin scattering," *Nat. Comm.*, vol. 8, p. 15819, 2017.
344. C. Rios, M. Stegmaier, Z. Cheng, N. Youngblood, C. D. Wright, W. H. P. Pernice, and H. Bhaskaran, "Controlled switching of phase-change materials by evanescent-field coupling in integrated photonics," *Opt. Mater. Express*, vol. 8, pp. 2455–2470, 2018.
345. Q. Zhang, Y. Zhang, J. Li, R. Soref, T. Gu, and J. Hu, "Broadband nonvolatile photonic switching based on optical phase change materials: beyond the classical figure-of-merit," *Opt. Lett.* vol. 43, pp. 94–97, 2018.
346. C. Wang, M. Zhang, X. Chen, M. Bertrand, A. Shams-Ansari, S. Chandrasekhar, P. Winzer, and M. Lončar, "Integrated lithium niobate electro-optic modulators operating at CMOS- compatible voltages," *Nature*, vol. 562, no. 7725, p. 101, 2018.
347. M. He, M. Xu, Y. Ren, J. Jian, Z. Ruan, Y. Xu, S. Gao, S. Sun, X. Wen, L. Zhou, L. Liu, C. Guo, H. Chen, S. Yu, L. Liu, and X. Cai, "High-performance hybrid silicon and lithium niobate Mach–Zehnder modulators for 100 Gbps and beyond," *Nature Photon.*, vol. 13, pp. 359–364, 2019.
348. S. Serna, H. Lin, C. Alonso-Ramos, A. Yadav, X. Le Roux, K. Richardson, R. Cassan, N. Dubreuil, J. Hu, and L. Vivien, "Nonlinear optical properties of integrated GeSbS chalcogenide waveguides," *Photon. Res.*, vol. 6, pp. B37–B42, 2018.
349. S. Richard, "Mid-infrared photonics in silicon and germanium," *Nature Photon.*, vol. 4, no. 8, pp. 495–497, 2010.
350. J. Wang, S. Paesani, Y. Ding, et al., "Multidimensional quantum entanglement with large-scale integrated optics," *Science*, vol. 360, no. 6386, pp. 285–291, 2018.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

30

Fibre-Optic Transmission Systems from Chip-to-Chip Interconnects to Trans-Oceanic Cables

Peter J. Winzer

CONTENTS

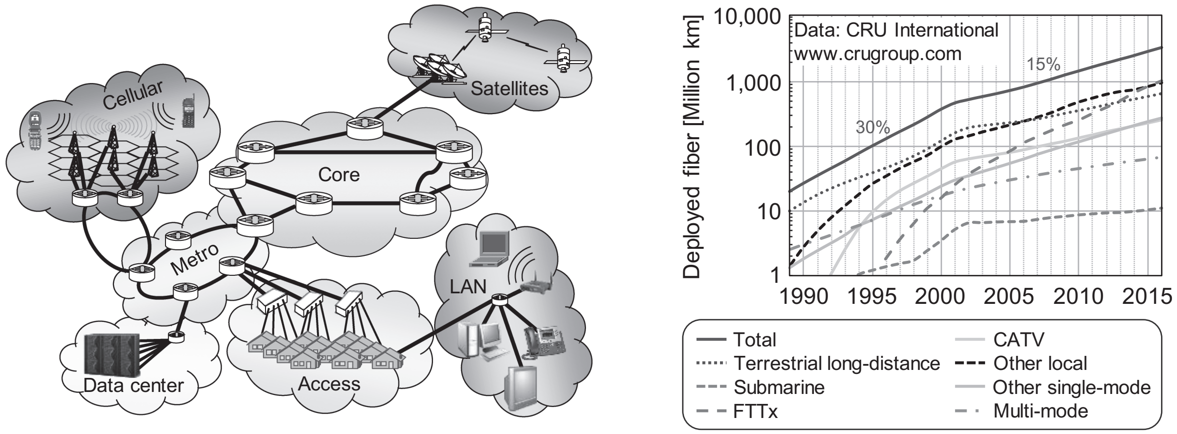
30.1 The Role of Fibre Optics in Communications 427
30.2 Modulation and Multiplexing: Theoretical and Practical Limits 428
  30.2.1 Digital Modulation and Multiplexing in a Nutshell 428
  30.2.2 Practical Aspects of Modulation and Multiplexing in Optical Communications 429
30.3 Fibre-Optic Communication Systems from Short-Reach to Ultra-Long-Haul 430
  30.3.1 Short-Reach Systems ($\lesssim$ 10 km) 430
  30.3.2 Mobile Front-Haul and PONs 432
  30.3.3 Datacentre Interconnects ($\lesssim$ 100 km) 433
  30.3.4 Metro and Long-Haul Networks (100 km ~ 5000 km) 434
  30.3.5 Sub-marine Transmission ($\gtrsim$ 5000 km) 439
References 440

30.1 The Role of Fibre Optics in Communications

Fibre-optic communication systems form the underlying technological basis of today's globally connected and highly automated communication and information society [1,2]: Commercially deployed since 1977 in terrestrial networks and since the late 1980s on trans-oceanic cables, fibre-optic communication systems were initially used predominantly for long-distance transmission, but have since also made their way into shorter reach applications, such as passive optical networks (PONs) for fibre-to-the-home (FTTH) access systems, systems for analog (and later digital) cable television distribution, cellular mobile fronthaul, as well as local-area and intra-datacentre networks, cf. Figure 30.1a; fibre-optics is currently taking its next step into opto-electronically integrated chip-to-chip interconnects. As of today, more than four billion kilometres of fibre are installed worldwide, enough to be wrapped around the globe approximately 100 000 times, cf. Figure 30.1b, and commercial optical fibre systems carry well over an Exabit/s (10^{18} bit s^{-1}) of traffic on a daily basis, growing at a rate of 45% per year, consistent for decades and driven by the exponential increase in computational power fuelled by Moore's law [2].

The enormous success of fibre-optic communications over the past four decades has two fundamental reasons:

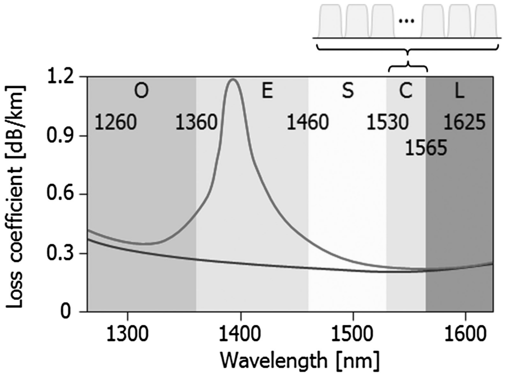
- Fused-silica optical fibre [3] has an exceptionally low loss coefficient of typically 0.2 dB km^{-1} (including deployment-related splices), with record demonstrations as low as 0.14 dB km^{-1} [4]. Figure 30.2 shows typical loss coefficients across the low-loss window of commercial single-mode optical fibre with (red) and without (blue) the characteristic hydroxyl group (OH) absorption peak around 1380 nm. The low-loss window is seen to extend over a range of more than 50 THz in the near infrared, from ~1260 to ~1625 nm, for which efficient system components (lasers, modulators, detectors, amplifiers) have also been developed. In addition to its low loss, optical fibre is thin, lightweight, flexible, easily spliceable, and cheap to manufacture (legacy fibre selling on the order of \$10 km^{-1}). As such, optical fibre has no competition from *any* other known guided-wave transmission medium.
- The increase in carrier frequencies f_c by between 4 and 7 orders of magnitude due to the transition from radio-frequency (RF) communications to optical communications in the 1970s afforded proportionally larger *absolute* system bandwidths B at the same *relative* system bandwidth $B_{rel} = B/f_c$. Keeping the latter fixed is important in engineering practice, as B_{rel} is a measure for the practical complexity of building a system [2]. (In many areas of engineering, systems typically operate at relative bandwidths below 10%. In the case of optical communications, for example, C-band Erbium-doped optical fibre amplifiers, EDFAs, cf. Figure 30.2a, have a relative amplification bandwidth of <3%, which at the high optical carrier frequency still represents almost 5



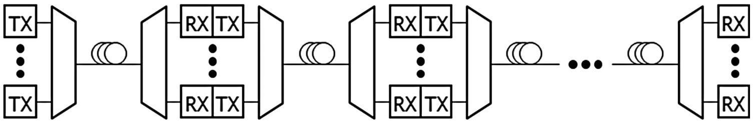
(a) Fiber-optics in communication networks

(b) Global fiber deployments

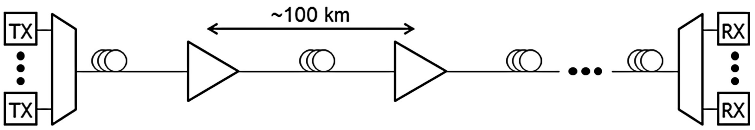
FIGURE 30.1 (a) The role of fibre optics in today's communication networks. (b) The globally deployed optical fibre infrastructure comprises more than 4 billion kilometres today and is growing at a cumulative rate of 15% per year (Figure after Ref. [2]).



(a) Attenuation coefficient of telecommunication fiber



(b) WDM system using per-wavelength opto-electronic regeneration



(c) WDM system using full-band optical amplification

FIGURE 30.2 (a) Optical fibre loss coefficient across the low-loss transmission window for commercial telecommunication fibre with and without the OH absorption peak. Typical telecommunication bands (O, E, S, C, L) are indicated. Within a band, WDM may be employed, as shown for the C-band. (b) WDM system using opto-electronic regeneration. (c) WDM system using optical amplification.

THz of absolute bandwidth.) The availability of efficient multi-channel optical amplifiers has enabled cost-effective long-haul fibre-optic system capacity scaling compared to earlier per-wavelength opto-electronically regenerated systems, Figure 30.2b and c. Larger absolute system bandwidths directly result in larger aggregate bit rates, both through the use of shorter optical pulses transmitted at higher rates (time-division multiplexing, TDM) and through the use of multiple parallel frequency channels (frequency-division multiplexing, FDM, which in fibre-optics is also referred to as wavelength-division multiplexing, WDM, cf. Figure 30.2). Note that the scaling from RF to optical carrier frequencies was done largely across the classical (thermal-noise-limited) frequency range (from the MHz – GHz range to ~ 6 THz) at normal ambient temperatures, and only in part (from ~6 to 193 THz) across quantum-noise-limited frequencies. This makes a further carrier frequency scaling, e.g. to extreme ultra-violet or X-ray frequencies much less attractive [5].

30.2 Modulation and Multiplexing: Theoretical and Practical Limits

30.2.1 Digital Modulation and Multiplexing in a Nutshell

Any system using electromagnetic waves (and as such also any fibre-optic communication system) operates within the constraints of exactly five physical dimensions available for *modulation* and *multiplexing* [6]. Viewing digital communications as the temporal concatenation of a series of letters (symbols) constructed from an *alphabet* agreed upon by a transmitter and a receiver [7], *modulation* refers to the formation of letters within the context of the available physical dimensions, while *multiplexing* refers to the parallel transmission of multiple symbol streams. The redundancy contained in words and sentences and the corresponding ability to detect typos and other linguistic errors are analogous to *forward error correction* (FEC) in digital communications. The five available physical dimensions are as follows:

- *Time* (i.e. the *symbol rate* and the employed analog *pulse shape*)
- *Quadrature* (i.e. the use of sine and cosine components of an optical wave)
- *Polarization* (i.e. the use of two (and only two) orthogonal polarization states)
- *Frequency* (i.e. the use of WDM across the system's bandwidth)
- *Space* (i.e. the use of parallel spatial paths or orthogonal spatial modes)

The maximum available information bit-rate that may be transmitted over a channel can be calculated using Shannon's famous *channel capacity* [8], which in its most well-known form applicable to a classical memoryless additive white Gaussian noise (AWGN) channel reads[1]

$$C = 2MB\log_2\left(1+\mathrm{SNR}\right) = 2MB\log_2\left(1+P/\left(2MBN_0\right)\right) \quad (30.1)$$

The channel's effective signal-to-noise ratio (SNR) is given by the ratio of the average overall signal power P (spread across 2 polarizations, M parallel spatial paths or, more generally, M orthogonal spatial modes, and a system bandwidth B) to the overall received noise power $2MBN_0$. In *power-limited* systems, such as short-reach chip-to-chip interconnects (Section 30.3.1) or sub-marine cables (Section 30.3.5), P is given by the available power for the entire system across all its degrees of freedom, while in *non-linearity-limited* systems, such as in most terrestrial fibre-optic applications (Section 30.3.4), P can be increased proportional to the number of multiplexing dimensions ($2MB$), limited by signal distortions from fibre non-linearities. In the context of equation (30.1), the noise power spectral density per polarization and spatial path, N_0, is modelled as signal-independent, additive, complex Gaussian, and white across the system's bandwidth, its two polarizations, and its spatial paths/modes (AWGN); N_0 captures a variety of physical noise sources, including non-linear distortions arising from Kerr non-linearity within the transmission fibre, cf. Section 30.3.4.

With the quantity

$$\mathrm{SE} = C/\left(2MB\right), \quad (30.2)$$

referred to as the *spectral efficiency* per spatial and polarization mode, i.e. the information capacity of a single-mode signal per unit bandwidth, equation (30.1) can be written as

$$\mathrm{SE} = C/\left(2MB\right) = \log_2\left(1+E_\mathrm{b}\ \mathrm{SE}/N_0\right), \quad (30.3)$$

with the energy per information bit E_b defined as

$$P = E_\mathrm{b}\ C = 2MBE_\mathrm{b}\ \mathrm{SE} \quad (30.4)$$

The goal of transmission system design is to maximize or minimize one of the parameters C, SE, P, E_b, B, or M while imposing constraints on some or all of the others. Note in this context that E_b is the received optical signal energy per *information bit*, not counting various energy overheads for signal amplification, conditioning, processing, or coding, nor any (in practice often dominating!) system inefficiencies [11]. Further boundary conditions are usually imposed by such aspects as cost-efficient manufacturability of the underlying high-speed opto-electronic hardware components and their integration, as well as digital signal processing (DSP) and coding complexity and latency, as we shall discuss in the context of various system application scenarios in Section 30.3.

30.2.2 Practical Aspects of Modulation and Multiplexing in Optical Communications

Up until the early 2000s, implementation-related boundary conditions typically outweighed more fundamental Shannon trade-offs in fibre-optic communications, and progress was mostly due to advances in optical, electronic, and opto-electronic device technologies, such as wide-band optical amplifiers, frequency-stable lasers, and narrow-band optical filtering components, which in the context of the above considerations allowed for larger system bandwidths B and larger SE. By the early 2000s, *lasers* had reached ~GHz frequency stabilities, *optical filters* had bandwidths allowing for ≤50-GHz WDM channel spacings, and electronically generated and transform-limited *modulation* spectra of 40-Gb s^{-1} binary optical signals filled these frequency slots. At this remarkable point in time where 'optical and electronic bandwidths had met', optical communications had to shift from a *device physics*-dominated field towards *communications engineering* to further increase transmission capacities. Consequently, by 2002, high-speed fibre-optic systems research had started to investigate advanced modulation formats [12]. At the same time, the capabilities of CMOS processing, fuelled by Moore's Law, and with it the speed of CMOS-integrated analog-to-digital converters (ADCs) and digital-to-analog converters (DACs) had caught up with then prevalent 10-Gb s^{-1} systems. The new CMOS capabilities were first used to equalize 10-Gb s^{-1} direct-detection systems, e.g. through maximum-likelihood sequence estimation (MLSE) [13] and through digital electronic dispersion pre-compensation [14]. These first attempts of *digitally processing* and *digitally synthesizing* optical waveforms culminated in the commercialization of digital coherent receivers in 2008 [15], with digitally synthesized optical Nyquist pulses used just a few years later. The ability to fully digitally synthesize and analyze the complete optical field

---

[1] A full quantum-mechanical version of Shannon's capacity formula was derived based on entropy arguments by Gordon in 1962 [9] and formally proven to be the capacity of an arbitrary quantum channel by Holevo in 1973 [10]; see Ref. [5] for a discussion of quantum capacities in the context of practical fibre-optic communications.

TABLE 30.1

Typical Short-Reach Technologies (Ethernet, IEEE 802.2)

Link Distance	Technology	Bit Rate Per Lane
$\lesssim 3$ m	Twin-axial electrical	$\lesssim 50$ Gb s^{-1}
$\lesssim 100$ m	850-nm VCSEL+MMF	$\lesssim 50$ Gb s^{-1}
$\lesssim 10$ km	1310-nm DFB+SMF	$\lesssim 50$ Gb s^{-1}

in amplitude, phase, polarization, and more recently also in its spatial-modal content is undoubtedly one of the greatest breakthroughs in optical communications over the past two decades, an ability with far-reaching potential in other areas of science and engineering, such as advanced digital holography with applications in, e.g. endoscopic imaging and optical tweezers.

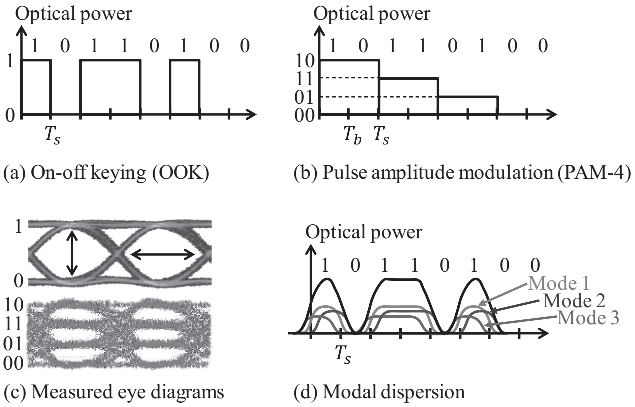
In the following chapter, we elucidate the diverse range of system applications, highlight their respective implementation constraints, and show which techniques and technologies are used to address unique system challenges in the light of the above, generally applicable discussion, with a perspective of both commercially available technologies and current research directions.

30.3 Fibre-Optic Communication Systems from Short-Reach to Ultra-Long-Haul

30.3.1 Short-Reach Systems ($\lesssim$ 10 km)

Commercial short-reach fibre-optic communications comprise transmission distances from a few metres to a few kilometres, predominantly with applications in local-area networks and intra-datacentre communications [16]. Connections up to a few metres with up to 50 Gb s^{-1} per lane using twin-axial *direct attach cables* are still being handled electrically today. However, neither much longer links at this speed nor much higher speeds over this distance are possible electrically, as cable bandwidths shrink considerably with cable length due to RF attenuation. This fundamental trade-off between bit rate and transmission distance, which has forced electrical technologies to ever shorter distances as bit rates have grown over the years, will continue to transform electrical interconnects to optical interconnects. Table 30.1 summarizes typical short-reach technologies in commercial use today and standardized in the framework of Ethernet (IEEE 802.3 [17]). As a next step, optics is expected to replace electronics for communication distances $\lesssim 10$ m, in particular, for high-throughput chip-to-chip interconnects such as high-end Ethernet switch chips, which are commercially available at electrical input/output (I/O) capacities exceeding 12 Tb s^{-1} today and are expected to reach 50-Tb s^{-1} capacities by 2022 [2]. In addition to the above bandwidth vs. distance trade-off, chip-to-chip interconnects will also be needed for interconnect density reasons [18].

In order to optically bridge distances up to ~100 m, systems typically use 850-nm Vertical Cavity Surface Emitting Lasers (VCSELs) coupled to multi-mode fibre (MMF). Simple on-off keying (OOK) at typically 10 or 25 Gb s^{-1} is used, i.e. the laser is switched between a low-power state, usually just above threshold, to represent a logical ‘0’ and a high-power state to represent a logical ‘1’, cf. Figure 30.3a. Some such systems are venturing to either 50-Gb s^{-1} OOK or to 25-GBaud 4-level pulse amplitude modulation (PAM-4), yielding a bit rate of 50 Gb s^{-1} as well. (The number of transmitted pulses per second is measured in units of *Baud*.) As shown in Figure 30.3b, four transmitted intensity levels convey 2 bits per symbol, which lets the symbol period T_s be twice the bit period T_b and the symbol rate $R_s = 1/T_s$ be half the bit rate $R_b = 1/T_b$. Labelling the four levels as 00/01/11/10 instead of the binary progression 00/01/10/11 is referred to as *Gray mapping* and avoids multi-bit errors in case the receiver mistakes an intensity level by its adjacent level due to thermal noise generated within the receiver, which is the limiting noise source in this kind of systems. Figure 30.3c shows measured eye diagrams at 53.5 Gb s^{-1} (OOK) and 28 GBaud (PAM-4), respectively. An *eye diagram* is obtained by recording an OOK or PAM waveform on an oscilloscope that is triggered at the symbol rate (or at a small integer fraction thereof). The horizontal and vertical *eye openings* give valuable information about the quality of the signal in terms of its level separation (vertical eye opening) and its sampling margin and timing jitter (horizontal eye opening).



(a) On-off keying (OOK)

(b) Pulse amplitude modulation (PAM-4)

(c) Measured eye diagrams

(d) Modal dispersion

FIGURE 30.3 (a) Binary OOK and (b) 4-level optical intensity modulation (PAM-4) as used for short-reach fibre-optic transmission systems; (c) measured 53.5-Gb s^{-1} OOK and 28-GBaud PAM-4 eye diagrams; (d) visualization of modal dispersion.

The reason for using simple intensity modulation of multi-mode VCSELs lies in the energy and cost-efficiency of VCSELs, and in particular in their straightforward and alignment-tolerant coupling to MMF, which significantly reduces (typically dominant) device-packaging costs. Even though MMF by itself is more expensive than single-mode fibre (SMF), the associated transponder packaging advantages make MMF-based interconnects more cost-effective for short-reach systems, which highlights an important general system design principle: *What matters is the cost of an overall system, not of any of its individual components.*

The multiple spatial modes generated by a VCSEL propagate at slightly different speeds within a MMF (*modal dispersion*), as visualized in Figure 30.3d. Since receivers in this application scenario are simple intensity detectors that only respond to the total optical power impinging onto them, independent of polarization, wavelength, or spatial-modal content of the received signal, modal dispersion, and random mode coupling within the MMF manifest in pulse broadening upon detection, which eventually closes the eye and results in detection errors. The effect of modal dispersion linearly increases with link distance and symbol rate, which lets the *bandwidth x distance product* be an appropriate metric for links limited by this effect. Commercial MMFs have bandwidth x distance products between ~ 200 MHz × km (optical-multimode-1, OM1) to ~ 3500 MHz × km (OM4 and OM5). Short-reach systems typically try to avoid FEC and, therefore, require physical-layer '*error-free*' operation (defined as having a bit error ratio below a value between 10^{-9} and 10^{-16}, depending on the system application). The omission of FEC reduces complexity, power consumption, and latency, which is particularly important for such highly integrated short-reach transponders.

Short-reach optical transponders are available either as *active optical cables,* i.e. pluggable transponders permanently connected to up to several ten metres of fibre, or as stand-alone tightly integrated *pluggable modules* of various physical form factors, cf. Figure 30.4. Short-reach pluggables with a reach of ~100 m have a power consumption of ~30 pJ bit^{-1} (30 mW Gbps^{-1}) and sell at ~1$ Gbps^{-1} today. As short-reach systems are entering the *chip-to-chip interconnect* space, both numbers will have to come down significantly, e.g. to below 1 pJ bit^{-1} and to below 10¢ Gbps^{-1}, in order to enable future ASICs with I/O capacities of 50 Tb s^{-1} and beyond.

Recent research on MMF-based short-reach transmission systems also investigates the use of *mode-group-multiplexing*, using mode-selective optical elements to launch different signals into different MMF mode groups and to separate mode groups before intensity detection at the receiver, thereby largely avoiding modal dispersion [19]. Such techniques could make economic sense whenever an expensive installed base of MMF must be re-used to upgrade systems to higher interface rates. (If new fibre deployments are an option, SMF fibre bundles are likely to be economically more attractive.)

For links between ~100 m and ~10 km, SMF is being used to avoid the limiting effect of modal dispersion at the required bit rates. The typical symbol rate for 100-Gb s^{-1} and 400-Gb s^{-1} applications is 25 GBaud and 50 GBaud, sometimes with an additional overhead for FEC, using either OOK or PAM-4, cf. the respective 100G and 400G IEEE 802.3 Ethernet standards [17]. The resulting power- and cost-efficient per-lane bit rates of 25 Gb s^{-1} to 100 Gb s^{-1} necessitate *parallelism* in the *spatial* (*M*) or *frequency* (*B*) domains, which for shorter link lengths (~500 m) can be done through four parallel strands of single-mode fibre (PSM-4) and for longer links is accomplished through parallel wavelengths (WDM with up to 8 wavelengths) [17]. Transponders of this kind are typically tightly integrated pluggable modules, as shown in Figure 30.4. While the power consumption remains at ~30 mW Gbps^{-1} for 10-km transponders, their cost can increase by a factor of ~10 compared to 100-m interfaces.

In terms of modulation technologies, and with reference to Figures 30.5a and b), transponders up to ~10-km reach at state-of-the-art bit rates typically use [12,21,22]:

a. *Directly modulated distributed feedback* (DFB) lasers (DMLs), i.e. single-frequency semiconductor lasers whose drive current $I = I_0 + I(t)$ is directly modulated by the data signal. At the high bit rates considered today, single-frequency laser operation is critical to avoid substantial chromatic dispersion (CD) penalties: Modulating a laser that oscillates at multiple longitudinal modes (as typical of Fabry-Perot lasers that are sometimes still used for bit rates below ~1 Gb s^{-1}) leads to multiple spectrally overlapping replica of the modulated signal that travel at slightly different speeds, as dictated by the fibre's CD (measured as a *ps*-delay per *nm* of frequency separation and per *km* of fibre propagation in units of [*ps (km nm)*$^{-1}$]). The superposition of these dephased spectral replica at the receiver results in a similar effect as shown for spatial-mode dispersion in Figure 30.3d. In addition, DMLs exhibit *chirp*, i.e.

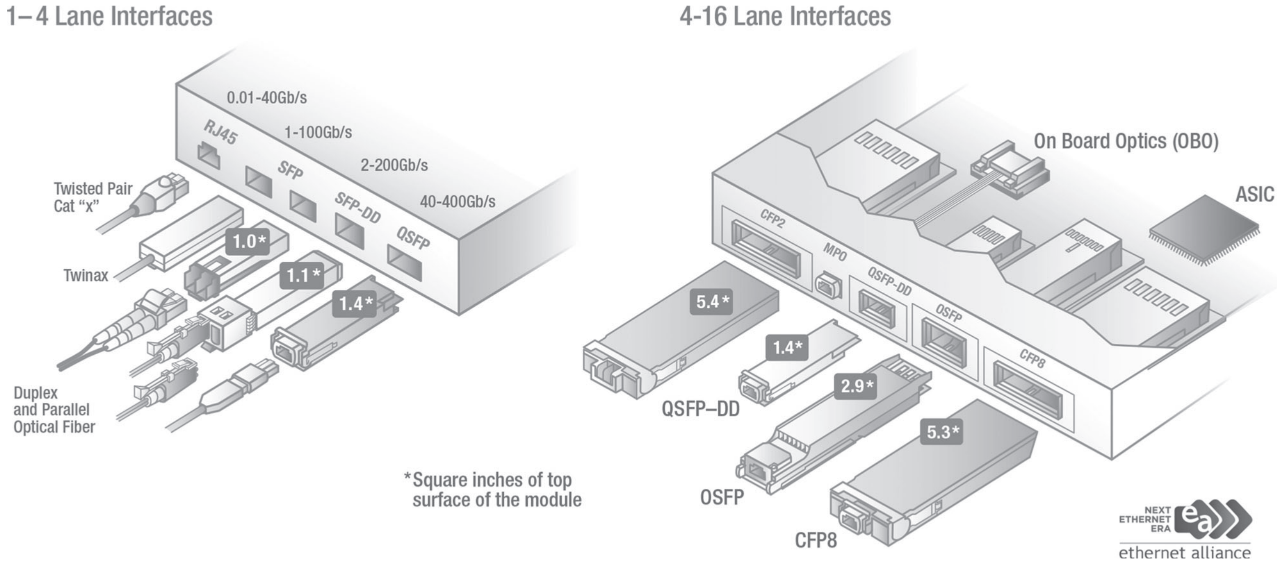


FIGURE 30.4 Typical form factors of 100G and 400G Ethernet pluggable modules [20].

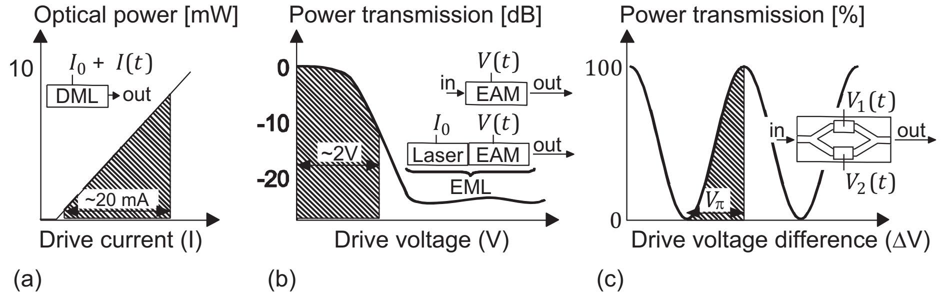


FIGURE 30.5 Short-reach transponders are typically based on (a) a directly modulated DFB laser (DML), an (b) EAM, sometimes integrated with a DFB laser as an EML; as reaches get longer (c) MZMs are being used.

residual frequency modulation accompanying the desired intensity modulation. The interplay between chirp and CD generally limits transmission distances significantly below those of un-chirped sources.

b. *Electro-absorption modulators* (EAMs), i.e. semiconductor *p-i-n* structures whose reverse-bias voltage *V*(*t*) controls the material's band edge and hence its absorption. EAMs are sometimes integrated with the DFB laser to form an *electro-absorption modulated laser* (EML). EMLs feature a more controlled chirp than DMLs that can even be tailored to partially counteract CD-induced pulse broadening [22], leading to longer transmission distances than those achievable with un-chirped signals.

For short-reach systems, where cost and power consumption are critically important design aspects, source lasers are often *uncooled*, with emission wavelengths varying within a 20-nm window across the operating temperature range, which allows for coarse WDM (CWDM) systems on a 20-nm wavelength grid. The preferred wavelength of operation is close to 1310 nm, where the low CD of standard SMF (SSMF) prevents detrimental optical pulse broadening. Transmitting at nominally 1310 nm and taking into account wavelength drifts of uncooled lasers as well as variations in the exact location of the zero-dispersion wavelength of SSMF, the practical CD-limited distance of an un-chirped 25-Gb s^{-1} OOK signal at 1310 nm is ~100 km. Note that in contrast to modal dispersion, the effect of CD scales *quadratically* with symbol rate, i.e. doubling the symbol rate shortens the CD-limited distance by a factor of four. This is because CD lets different frequencies contained within the modulated optical spectrum travel at different speeds, resulting in propagation delay differences that increase linearly with the signal's bandwidth (and hence linearly with the symbol rate). In addition, the effect of a given amount of dispersive pulse broadening becomes *relatively stronger* the shorter the symbols are (e.g. a 10-ps pulse spreading corresponds to 10% of the symbol duration at 10 GBaud but to 50% at 50 GBaud). The combination of these two linear effects results in an overall *quadratic* scaling of the CD-limited reach with symbol rate. The CD-limited distance for un-chirped OOK signals is ~10 km at 25 Gb s^{-1} over SSMF at 1550 nm [~17 ps (km nm)$^{-1}$].

30.3.2 Mobile Front-Haul and PONs

Two important applications of ~10-km interconnect technologies are *mobile fronthaul* and *PONs*. While the general physical-layer optical transmission technologies are essentially those described in Section 30.3.1, we will point here to some specific features characterizing these two important applications.

As shown in Figure 30.6a, *mobile fronthaul* [23,24] is used to separate (i) the analog RF transmit/receive functionality including RF mixers, amplifiers, and antennas (the remote radio head, RRH) for cellular telephone and data services and (ii) the required underlying DSP functionality (the baseband unit, BBU), as it is desirable for size, weight, power, operating temperature, and security reasons to place as few components onto a cell tower as possible. The dominant protocol used for the purpose of connecting RRH and BBU is the *Common Public Radio Interface* (CPRI) standard, initially used to replace coaxial RF cables and connect a BBU to its associated RRH by much lighter and longer-reaching optical fibres, which provided more flexibility in the placement of the BBU near its associated cell tower. Once inside a fibre, signals can propagate over long distances with little loss, which further enables the pooling of multiple cell sites into a common BBU site, known as a *centralized (or cloud) radio access network* (C-RAN). This allows for further economic synergies through site consolidation and also offers technologic benefits from advanced wireless interference mitigation by co-processing signals from multiple cells. Since the RRH may only demodulate to baseband and digitize (typically at ~12 bits per sample) the RF signal but may not process it much further, CPRI data rates are much higher than the actual mobile user data rates, which requires tens to hundreds of Gb s^{-1} in support of a cell tower. In addition, faithful handling of the sampled RF signals comes with stringent latency and jitter requirements, which results in a maximum fronthaul distance of $\lesssim$ 20 km and lets such aspects as wavelength walk-off between coarsely spaced WDM channels becomes a potential area of concern.

PONs [25,26], as shown in Figure 30.6b, again make use of the general optical transmission technologies discussed in Section 30.3.1. Apart from the most stringent cost constrains among all optical transmission applications, the main physical-layer differences of PON transponders compared to the short-reach solutions of Section 30.3.1 are (i) the very high losses that have to be accommodated from transmitter to

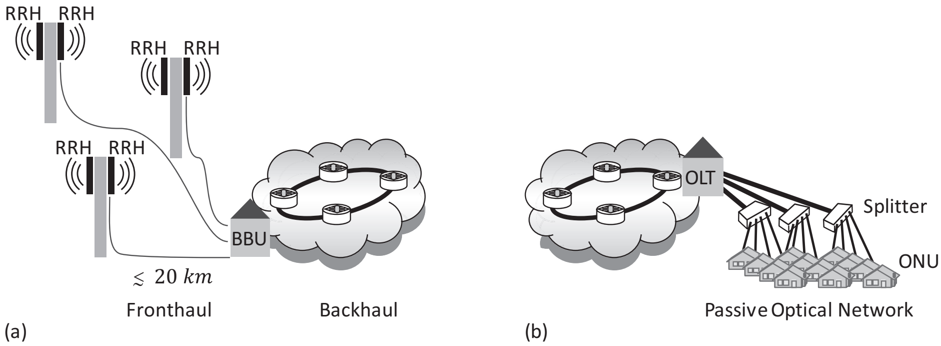


FIGURE 30.6 (a) General mobile fronthaul architecture used to optically connect remote radio heads (RRHs) and baseband processing units (BBUs). (b) General PON architecture connecting an OLT with many optical networking units (ONUs) on a fibre-tree topology.

receiver (as much as 30 dB) and (ii) the required burst-mode capabilities. Both are rooted in the fact that PONs are ultra-low-cost access networks built as a single-fibre tree based on passive optical power splitters between the optical line terminal (OLT) and the optical network unit (ONU) located, e.g. at a customer's home (FTTH). The distance between OLT and ONUs can be several 10 km, and in long-reach PONs even on the order of 100 km. Split ratios of 1:32 and higher are typical, depending on the available optical power budget. The higher the split ratio, the more users can share a common fibre tree and the more cost-effective a solution becomes to the operator. In contrast to the solutions discussed in Section 30.3.1, which typically have a separate fibre for each transmission direction, the single-fibre tree used in PONs requires two different wavelengths, one for downstream (from OLT to ONUs, at 1490 nm) and one for upstream (from ONUs to OLT, at 1310 nm to limit the effects of CD for uncooled DMLs). As all users on a PON share the same single-fibre tree, and user signals are only differentiated by their allocated time slots (TDM-PON), careful time scheduling must be used to ensure that no up-stream packets generated by the users' ONUs collide upon arrival at the OLT. The resulting need for burst-mode transmission and burst-mode detection of the physical-layer hardware has implications on timing recovery as well as on laser wavelength stability. Various evolution scenarios of PONs also consider the wavelength domain as a means to multiplex users onto a common fibre tree (WDM-PON), which comes at the cost of wavelength splitters instead of power splitters as well as the need for more precisely wavelength-controlled low-cost lasers. Coherent detection (coh) is also being discussed for PONs [26,27] but has not been commercially implemented due to the lack of sufficiently low-cost components; PON transponders represent one of the most *cost-constrained* class of optical transponders, and as such, any innovation in this space must first and foremost consider ONU costs. (OLT costs are usually of slightly lesser concern, as OLTs reside in the operator's central office and are shared by multiple users.)

30.3.3 Datacentre Interconnects ($\lesssim$ 100 km)

As transmission distances get longer, the impact of the transmission fibre on system design becomes more pronounced, both in terms of *performance* and in terms of *cost*.

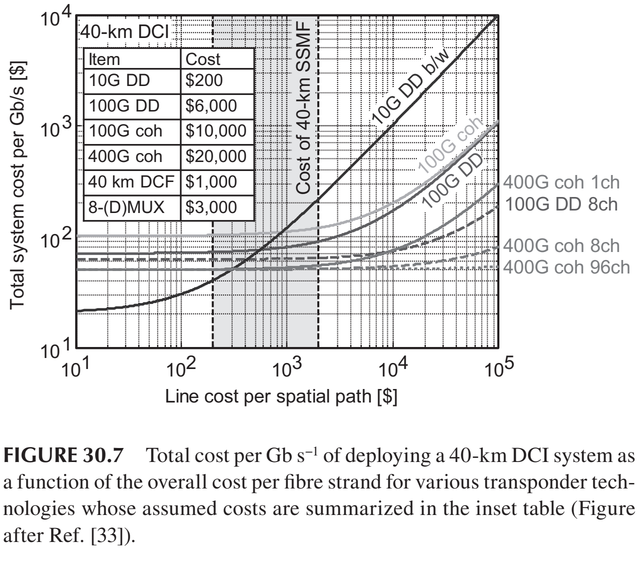
In terms of *performance*, reaching ~100 km at 100 Gb s^{-1} and above requires advanced modulation and/or advanced detection techniques, first and foremost to combat limitations by CD. As the CD-limited range of a practical low-cost (uncooled) 25-GBaud transmitter at 1310 nm is ~100 km (cf. Section 30.3.1), multi-level modulation (e.g. PAM-4 [28]) enables 50 Gb s^{-1} per wavelength. The required multiplexing of a 100-Gb s^{-1} bit stream onto two CWDM wavelengths reduces the CD-limited transmission distance to ~50 km, and the reach of interfaces using 8×50 Gb s^{-1}=400 Gb s^{-1} becomes even more severely limited. The need for dispersion-compensating fibre (DCF) as part of the link as well as the associated need for optical amplification of more densely spaced wavelengths leads to such systems operating at 1550 nm [28]. Lasers used in such applications are either (inherently fixed-wavelength) DFBs or external-cavity tunable single-frequency lasers, the latter allowing for wavelength-division multiplexing based on software-defined wavelength selection. Modulation in the state-of-the-art systems for applications beyond several 10 km is typically based on *Mach-Zehnder modulators* (MZMs), cf. Figure 30.5c, which split the incoming laser light into two arms, phase-modulate one or both arms, and re-combine the phase-modulated light to generate constructive or destructive interference at the MZM output,

$$\begin{aligned} E_{\text{out}} &= \frac{1}{2}E_{\text{in}}\left(\mathrm{e}^{\frac{j\pi V_1(t)}{V_\pi}} + \mathrm{e}^{\frac{j\pi V_2(t)}{V_\pi}}\right) \\ &= E_{\text{in}}\mathrm{e}^{j\pi(V_1(t)+V_2(t))/2V_\pi}\cos\left(\pi\left(V_1(t)-V_2(t)\right)/2V_\pi\right) \end{aligned} \tag{30.5}$$

In order to generate chirp-free (pure amplitude) modulation of the optical field, push-pull operation is being used, i.e. $V_1(t) = -V_2(t)$, leading to the cancellation of the phase term on the right-hand side of equation (30.5) and thus in a transform-limited optical signal that suffers minimally from CD-induced pulse broadening [12]. The voltage difference between the two arms that is needed to exert a π-phase shift, switching the modulator output from maximum transmission to maximum extinction, is referred to as the *switching voltage* V_π. MZMs are available in various material systems, most notably lithium niobate (LiNbO$_3$), indium phosphide (InP), and more recently also in silicon photonics (SiPh), and represent the most ideal-behaving and most versatile, albeit not the most compact modulator type in use today.

Alternatively, advanced detection techniques such as (commercial) coherent receivers allow for the digital compensation of CD without the need for DCFs (cf. Section 30.3.4). As coherent transponders are significantly more complex than PAM-4 transponders, potentially more cost-effective self-coherent approaches that also allow for digital CD compensation are being heavily investigated in research [29–31]. However, as optical integration continues to progress [32], the enormous flexibility offered by digital coherent transponders together with the associated production volume is expected to eventually make this the dominant solution in the DCI market.

To assess the role of *fibre costs* in a 40-km DCI systems context, Figure 30.7 shows the total system cost per Gb s^{-1} as a function of the cost per fibre strand. The latter can vary widely in practice, as the cost of new fibre in a systems context is dominated by the cost of the cable deployment and not by the cost of the bare fibre itself. For example, if ample parallel strands of fibre are already deployed between two datacentre sites (often several 10,000 fibre strands linking datacentre buildings) the cost of fibre is essentially zero to the transmission system designer as long as the resulting system stays below the deployed fibre count. In cases where new cable deployments are needed, the incremental cost of an additional fibre within a new cable is small by itself, which drives operators to installing high-fibre-count cables if new deployments are unavoidable. If the deployment of new cables requires the construction of new ducts in a metropolitan area, deployment costs may be prohibitively high. Figure 30.7 compares transponders at 10 Gb s^{-1}, 100 Gb s^{-1}, and 400 Gb s^{-1} using direct detection (DD) and coh. All necessary line system components such as DCFs, WDM (de)multiplexers and amplifiers are included. If fibre costs are negligible, the system cost depends exclusively on the transponder cost per b s^{-1}. With the transponder cost assumptions made here, the lowest-cost system uses DD 10-Gb s^{-1} single-wavelength (black & white, b/w) interfaces, albeit at massive requirements on the available parallel fibre infrastructure. (For example, a 50-Tb s^{-1} DCI system built this way requires 10 000 parallel fibre strands.) Going to a 100-Gb s^{-1} DD solution (e.g. using 25-GBaud dual-wavelength PAM-4 and then-needed DCF, solid red in Figure 30.7) increases system costs relative to 10-Gb/s if fibre costs are negligible. If fibre costs are significant, the 100-Gb s^{-1}-based system becomes cheaper than the 10-Gb s^{-1} solution. The cost range of an additional fibre strand (provided that cabling and installation are absorbed elsewhere) is shown as a shaded region in Figure 30.7. Using moderate (e.g. 8-channel) WDM, the cost of the system is further reduced. Coherent transponders at 100 Gb s^{-1} are not cost-advantageous in the assumed cost model (green curve in Figure 30.7), but packing more bits onto a laser carrier amortizes coherent transponder costs across more transmitted information bits, making 400-Gb s^{-1} coherent transponders an attractive solution (solid blue). Depending on the cost per spatial path, the additional cost reduction of WDM can be very limited, though, and massive (96-channel) WDM systems only prove at very high fibre costs (dashed and dotted blue curves).



Item	Cost
10G DD	$200
100G DD	$6,000
100G coh	$10,000
400G coh	$20,000
40 km DCF	$1,000
8-(D)MUX	$3,000

FIGURE 30.7 Total cost per Gb s^{-1} of deploying a 40-km DCI system as a function of the overall cost per fibre strand for various transponder technologies whose assumed costs are summarized in the inset table (Figure after Ref. [33]).

30.3.4 Metro and Long-Haul Networks (100 km ~ 5000 km)

In contrast to short-reach optical transmission systems, which are heavily constrained by power, footprint, and cost considerations, and DCI systems, which are constrained by a trade-off between transponder and fibre costs, terrestrial metropolitan-area (‘metro’) and long-haul fibre-optic networks are typically *fibre-constrained*, as the cost to deploy new fibre is often prohibitively expensive, at ~$20 000 km^{-1} assuming already available duct space, and much more if new ducts need to be constructed. Deploying new long-haul routes is expensive because of the sheer lengths involved (e.g. across an entire continent) while deploying new metro fibre is expensive because of right-of-way legalities and the general difficulties associated with major construction projects within big cities. These costs place metro and long-haul systems off the scale to the right in Figure 30.7, resulting in the desire to deploy high-capacity WDM systems at the highest possible capacity, such as to utilize the existing fibre infrastructure as much as technologically possible. Figure 30.8 shows the evolution of commercially available WDM system capacities (triangles) as well as per-wavelength interface rates (circles) and symbol rates in CMOS-integrated coh systems (squares) over the past decades. While WDM systems of the 1990s were able to double their capacity each year, largely based on the fact that the useable bandwidth of optical amplifiers saw significant increases in these early days of erbium-doped amplifiers, the growth in WDM capacity slowed to ~20% in the early 2000s, enabled by the move from direct (optical intensity) detection to coherent (optical field) detection that allowed the use of (i) both signal polarizations (polarization-division multiplexing, PDM) as well as (ii) the modulation of both real and imaginary parts (also known as in-phase (I) and quadrature (Q) components) of the optical field for information transport. This inherent quadrupling of the degrees-of-freedom (2 polarizations × 2 quadratures) has been key to scaling WDM capacities since ~2010.

Some representative *quadrature amplitude modulation* (QAM) formats resulting from I/Q modulation are shown in

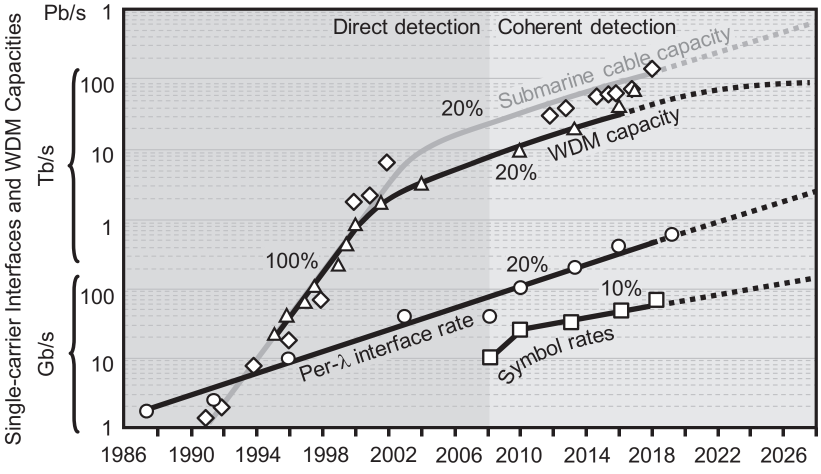


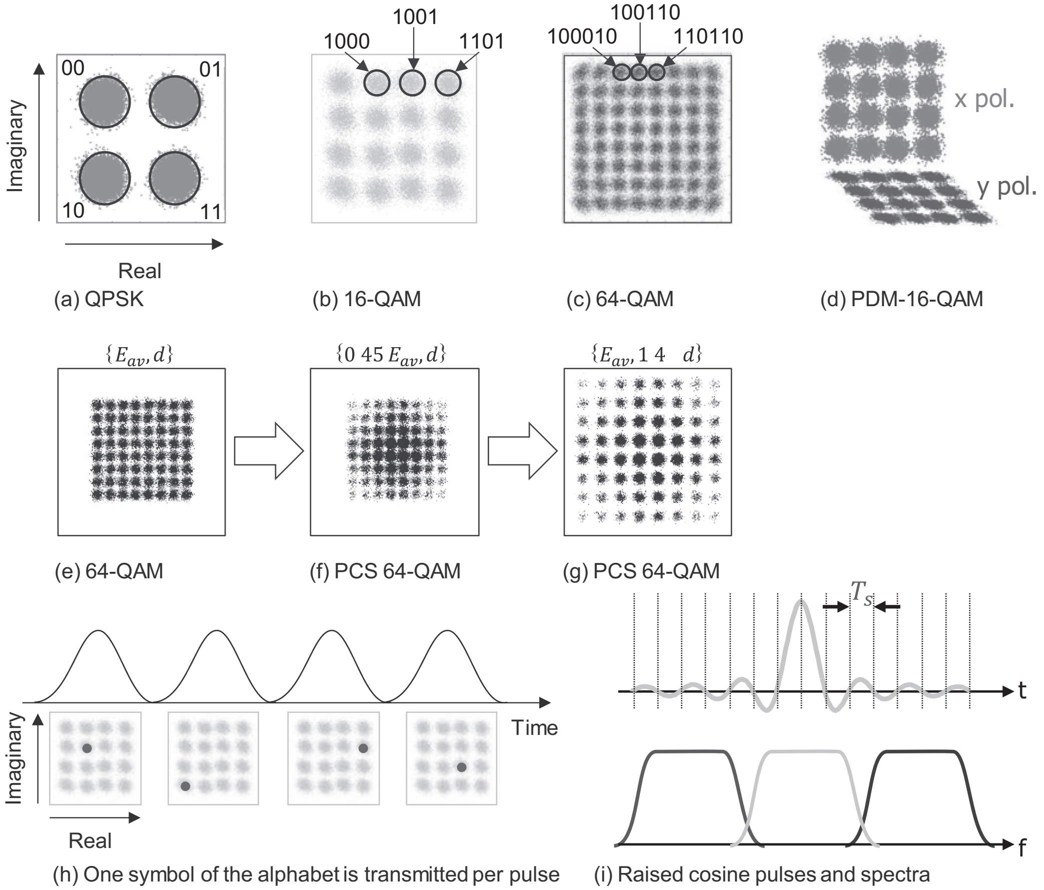
FIGURE 30.8 Evolution of commercial optical transmission technologies in terms of achievable per-wavelength interface rate (circles) and per-fibre WDM capacities (triangles). Starting in 2008, fibre-optic transport transitioned from DD to coh. The evolution of CMOS-integrated coherent symbol rates is shown as square markers. The evolution of sub-marine cable capacities is captured by diamond markers.

Figure 30.9a–c [7]. Quadrature phase-shift keying uses only two amplitude levels per quadrature, resulting in an alphabet of 4 complex values, each representing $\log_2 4 = 2$ bits of information (assuming equal probabilities of sending each complex value); similarly, 16-QAM conveys $\log_2 16 = 4$ bits per symbol and 64-QAM $\log_2 64 = 6$ bits per symbol. As the constellation order increases, the resilience to noise is reduced, since the distance d between constellation points shrinks for a fixed average transmit energy E_{av} per symbol, reflecting the general trade-off between SNR and capacity discussed along with equation (30.1). Polarization multiplexing doubles the number of bits transmitted per symbol by sending two independent signal streams in parallel, one per polarization, cf. Figure 30.9d. In addition to uniform square QAM, the latest generation of coherent transponders uses *probabilistic constellation shaping* (PCS) [34,35], whereby symbols are no longer transmitted at equal probabilities. Instead, sending the higher energy symbols towards the corners of a constellation less frequently than the lower energy symbols towards the centre reduces the average transmit energy, e.g. by 55% in the example shown in Figure 30.9e and f. Conversely, for the same average transmit energy, PCS allows for a larger symbol spacing and is hence more resilient to noise compared to uniform QAM, cf. Figure 30.9g.

In order to carry the complex values of the underlying QAM alphabets, optical pulse trains at a repetition rate of the symbol rate (typically between 30 and 70 GBaud, cf. square markers in Figure 30.8) are being used, whereby each pulse is multiplied by a complex number representing one letter of the symbol alphabet, as sketched in Figure 30.9h. However, the use of short pulses as schematically shown in Figure 30.9h would be very wasteful in terms of their modulation spectrum, which is inversely proportional to the pulse width. Maximum spectral compactness (and hence the densest possible WDM channel packing) without introducing temporal interference between adjacent pulses at integer multiples of the symbol period is achieved by *raised-cosine* pulses, Figure 30.9i. (Strictly speaking, transmitters use square-root-raised-cosine pulses, which become raised-cosine after matched filtering at the receiver [7].) Such pulses, including the multiplication with the complex symbols, are generated within a transmit-side DSP unit, followed by a DAC that generates the associated analog voltage waveforms driving an optical modulator. State-of-the-art CMOS-integrated DACs have sampling rates as high as 120 GSamples s^{-1} with 8 bits of resolution and must be integrated with the CMOS-DSP to accommodate the large internal bit rates. (For example, 4 DACs plus 4 ADCs, as needed in a single-channel coherent transponder, generate an aggregate bit rate of 8 converters × 8 bits resolution × 120 GSamples s^{-1} = 7.7 Tb s^{-1} to be passed between the converters and the DSP.)

As shown in Figure 30.10a, four electrical signals are needed to generate an optical PDM-QAM signal (I and Q components of x and y polarization). For each polarization, an I/Q-modulator is used as shown in Figure 30.10b. The modulator is based on two amplitude-modulating MZMs, cf. Figure 30.5c, which are nested inside a super-Mach-Zehnder structure that establishes the required 90-degree phase shift between the two quadrature components by means of a half-cycle optical delay in one of its arms, which is fine-tuned by a 90-degree phase shifter (V_{90}) [12]. The set-up is duplicated for PDM, with a polarizing beam splitter combining the x- and y-modulated transmit signals. Fully integrated PDM-I/Q-modulators are available in various material systems today, including LiNbO$_3$, InP, and SiPh [32].

The corresponding *coherent receiver* [36,37], shown in Figure 30.10c, splits the incoming signal into the *receiver's x/y* polarization reference system, which is different from the *transmitter's x/y* polarization reference system due to random polarization rotations within the transmission fibre, described by the channel's Jones matrix [21]. Hence, each polarization component at the receiver is initially a linear combination of the transmitter's x and y signals. Full-field coh is then based on the fact that the intensity of the combination of a signal optical field and a reference local oscillator (LO) laser optical field, if aligned in its polarization and matched in its spatial mode (of which there is only one in a SMF), contains a beat term that is linearly proportional to the signal optical field's complex amplitude,



(a) QPSK (b) 16-QAM (c) 64-QAM (d) PDM-16-QAM

(e) 64-QAM (f) PCS 64-QAM (g) PCS 64-QAM

(h) One symbol of the alphabet is transmitted per pulse (i) Raised cosine pulses and spectra

FIGURE 30.9 Coherent systems use QAM in conjunction with PDM to quadruple the bit-rate compared to direct-detection systems (a–d). Advanced transponders employ PCS (e–g) and digitally shape the optical spectrum to achieve maximally compact raised-cosine pulse shapes (i).

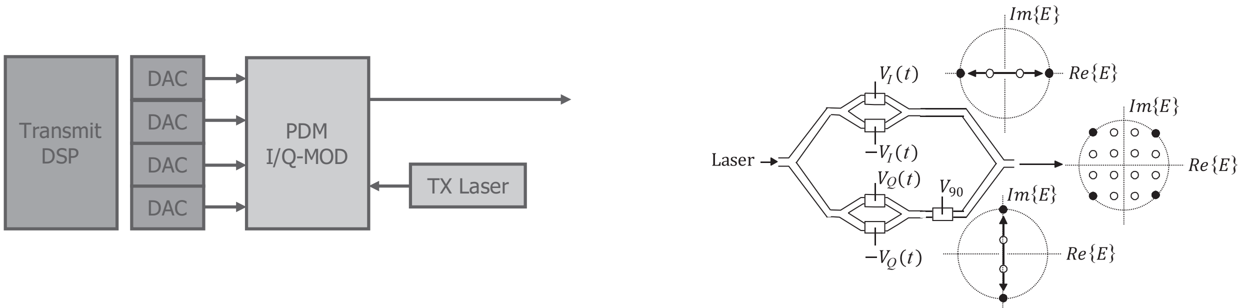
$$\begin{aligned} I_\pm(t) &= \left|E_{\mathrm{sig}} \pm E_{\mathrm{LO}}\right|^2 \\ &= E_{\mathrm{sig}}^2 + E_{\mathrm{LO}}^2 \pm 2Re\left\{E_{\mathrm{sig}}E_{\mathrm{LO}}^*\right\} \\ &= P_{\mathrm{sig}} + P_{\mathrm{LO}} \pm 2Re\left\{\sqrt{P_{\mathrm{sig}}P_{\mathrm{LO}}}\,\mathrm{e}^{j2\pi\Delta ft}\,\mathrm{e}^{j2\pi\left(\varphi_{\mathrm{sig}}-\varphi_{\mathrm{LO}}\right)}\right\}. \end{aligned} \tag{30.6}$$

The '±' sign reflects the fact that a lossless beam splitter fundamentally introduces a 180-degree phase shift between its two combined input signals at its two output ports [21] and allows for *balanced detection* to suppresses common-mode intensity terms arising from the signal-LO superposition process,

$$\Delta I = I_+ - I_- = 4Re\left\{\sqrt{P_{\mathrm{sig}}P_{\mathrm{LO}}}\,\mathrm{e}^{j2\pi\Delta ft}\,\mathrm{e}^{j2\pi\left(\varphi_{\mathrm{sig}}-\varphi_{\mathrm{LO}}\right)}\right\}. \tag{30.7}$$

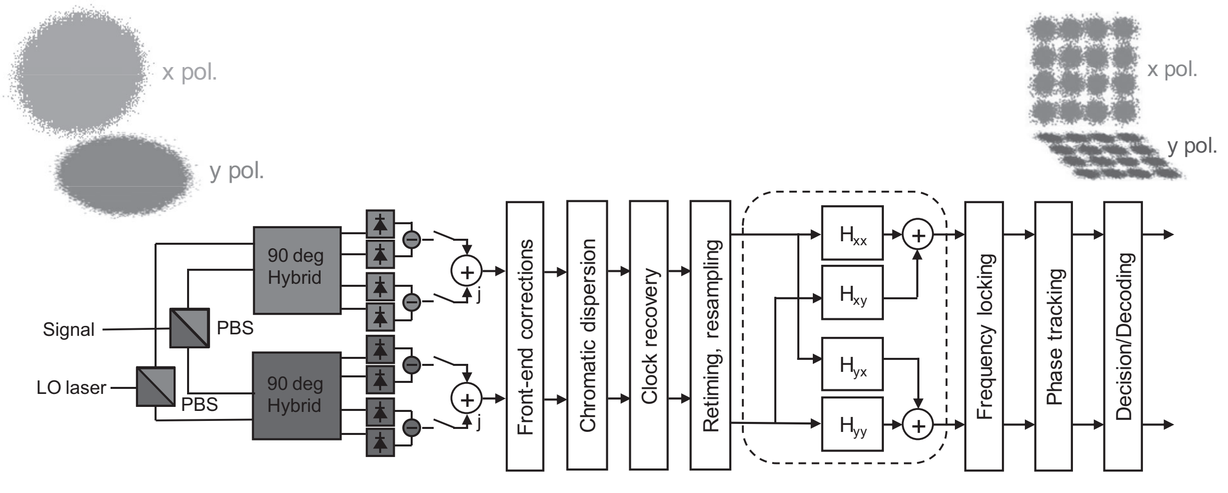
The beat signal amplitude oscillates at the difference frequency Δf between the signal's and the LO's laser carrier frequencies and exhibits a phase shift reflecting their phase difference $\varphi_{\mathrm{sig}} - \varphi_{\mathrm{LO}}$. In order to access the imaginary part of the optical field, the signal is split at the receiver's input and detected separately with two 90-degree-shifted copies of the LO, an operation performed by a *90-degree hybrid*. The two photocurrents at the output of a 90-degree hybrid with balanced detection represent the signal optical field referenced to the LO frequency and phase in one polarization, and the polarization-diversity set-up of Figure 30.10c gives access to both polarizations. In a digital coherent (intra-dyne) receiver, the LO wavelength is roughly aligned with the signal (typically to within a few GHz) but is neither frequency nor phase-locked with the transmit laser. Integrated modules that contain both transmit and coherent receive opto-electronics are commercially available today [32]. After conversion to the analog electrical domain, the four analog electrical outputs of the coherent receiver are sampled using four fast ADCs, which makes the complete received optical field accessible to receive-side DSP in powerful CMOS ASICs. Since the entire optical field is fully available in digital form at that point, frequency recovery and phase-locking can be done all-digitally, and arbitrary amounts of CD as well as other optical impairments can be fully compensated for by digital equalization, including the estimation and subsequent inversion of the Jones matrix representing the random transmit fibre polarization rotations [37,38].

Using the above modulation and detection techniques, SE vs. distance trade-offs shown in Figure 30.11 have been experimentally achieved. Note that the *y*-axis may represent (i) the achievable SE (as measured in the respective experiments), (ii) the achievable system capacity within a given


Transmit DSP
DAC
DAC
DAC
DAC
PDM I/Q-MOD
TX Laser
$Im\{E\}$
$V_I(t)$
$Re\{E\}$
$Im\{E\}$
Laser
$-V_I(t)$
$V_Q(t)$
V_{90}
$Re\{E\}$
$Im\{E\}$
$-V_Q(t)$
$Re\{E\}$

(a) PDM-QAM transmitter structure

(b) I/Q modulator (for one polarization)


x pol.
y pol.
x pol.
y pol.
90 deg Hybrid
90 deg Hybrid
Signal
PBS
LO laser
PBS
j
j
Front-end corrections
Chromatic dispersion
Clock recovery
Retiming, resampling
H_{xx}
H_{xy}
H_{yx}
H_{yy}
Frequency locking
Phase tracking
Decision/Decoding

(c) Digital coherent receiver

FIGURE 30.10 (a) Schematic of a coherent PDM-QAM transmitter based on (b) a double-nested MZM (I/Q-modulator). (c) Schematic of a digital coherent (intra-dyne) receiver including optical front-end and DSP back-end.

system bandwidth such as the C-band ($C = \mathrm{SE} \times B$), or (iii) an estimate for the achievable per-wavelength interface rate at a certain symbol rate $\left(R_\mathrm{b} \approx \mathrm{SE}\ R_\mathrm{s}\right)$. The shaded area indicates the region beyond the Shannon limit for standard single-mode fibre (SSMF), which is inaccessible to any classical communication scheme and has been estimated in [39–43]. Significant *transponder implementation penalties* prevent higher order formats at shorter distances to approach this limit as closely as lower order formats permit this at longer distances. In any case, we note that the experimental records shown in Figure 30.11 only have fairly modest gaps to the Shannon limit (~7 dB in terms of distance or SNR, or a ~40% headroom in terms of capacity at ~1000 km reach), which lets system engineering approach a point of diminishing returns (an '*economic Shannon limit*').

The Shannon limit is obtained from the *effective SNR* of the channel, including a host of noise sources, the most important of which are [42]:

- Optical amplifier noise (amplified spontaneous emission, ASE) arising from the periodic amplification of the signal, cf. Figure 30.2c, and leading to AWGN at the receiver. In a transmission system with N identical fibre spans of dB-loss ℓ_dB followed by lumped amplification of dB-gain $G_\mathrm{dB} = \ell_\mathrm{dB}$ and noise figure NF, the optical SNR (OSNR), defined as the ratio of received signal power to dual-polarization ASE power in a fixed 12.5-GHz reference bandwidth, is given by [42]

$$\mathrm{OSNR}_\mathrm{dB} = 58 + P_\mathrm{dB} - \mathrm{NF} - \ell_\mathrm{dB} - 10\log_{10} N \qquad (30.8)$$

More sophisticated optical amplification techniques relevant to long-haul optical transmission systems are

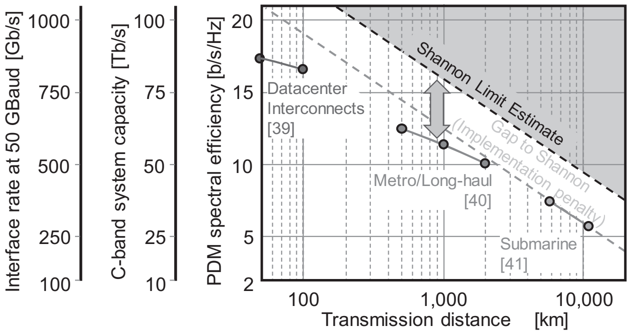


FIGURE 30.11 Experimentally achieved record SEs (in two polarizations) versus transmission distance (markers), shown with the Shannon limit estimate on SSMF (dashed). Two alternate interpretations of the y-axis represent WDM capacity and per-wavelength interface rate (Figure after Ref. [2,33]).

discussed in [42,44]. In a PDM system, the OSNR is related to the SNR (which is typically defined as signal power to noise power within a signal bandwidth corresponding to the modulation symbol rate) as [42]

$$\mathrm{SNR}_{\mathrm{dB}} = \mathrm{OSNR}_{\mathrm{dB}} - 10\log_{10} R_s/12.5\ \mathrm{GHz}. \qquad (30.9)$$

Importantly, the '*delivered SNR*' obtained this way characterizes only the amplification aspect of a line system, assuming a perfect transmitter and receiver and no further transmission-related degradations. In reality, the *'effective SNR'* to be used in a Shannon context to obtain capacity and performance is degraded by various transmitter, receiver, and system imperfections.

- The most important degradations in a long-haul fibre-optic systems context are non-linear signal distortions arising from fibre Kerr non-linearity in combination with CD, both acting within a WDM channel (*intra-channel non-linearities*) and between WDM channels (*inter-channel non-linearities*), and are described by the non-linear Schrödinger equation [42,45]. In the context of modern, dispersion-uncompensated coherent transmission systems, the aggregate effect of both types of fibre non-linear distortions can be accurately modelled as an equivalent AWGN source called *non-linear interference noise* (NLIN) [46,47]. The effects of NLIN fundamentally distinguish the fibre-optic channel from almost any other communication channel. Most importantly, NLIN leads to an *effective peak SNR*, representing the maximum signal power to be launched into a fibre span to achieve the highest OSNR while not experiencing too severe degradations from fibre non-linearity. A typical performance curve is shown in Figure 30.12a. Various techniques to mitigate NLIN aim at shifting the curve to higher launch powers, hence allowing for higher peak SNRs [48].
- In an optical mesh network context, cf. Figure 30.12b, WDM channels are frequently re-routed within the network using reconfigurable optical add/drop multiplexers [49,50]. Imperfections in these devices result in crosstalk, classified as either in-band crosstalk and out-of-band crosstalk. In the former, the crosstalk signal spectrally overlaps with the wanted signal, which typically results in more severe performance degradations than what is found in the case of spectrally disjoint crosstalk.

The trade-off between SE and reach is seen to follow a *logarithmic* relationship, with each doubling in transmission reach reducing the dual-polarization SE by ~2 bits/s/Hz, owing to the fact that the accumulated ASE scales linearly with the number of spans (or with transmission distance in the case of uniform amplification) and the SNR at the optimum signal launch power (including NLIN) scales to an excellent approximation inversely proportional to L. Still, optical transmission records are sometimes quoted in terms of their 'capacity×distance' product ($C\times L$) or for a fixed system bandwidth B in terms of their 'SE×distance' product ($\mathrm{SE}\times L$). It is important to note, though, that these product metrics date back to early multi-mode transmission systems limited by modal dispersion, whose impact on transmission system performance is to first-order proportional to the signal bandwidth (cf. Section 30.3.1), but are *inadequate* for modern single-mode fibre-optic transmission systems, which are limited by Shannon's capacity, equation (30.1). The correct system performance metric for such systems is given by the product metric

$$p = 2^{\mathrm{SE}} \times L = 2^{\mathrm{SE}_2/2} \times L. \qquad (30.10)$$

Hence, it is much harder to double SE than to double L in a modern fibre-optics communication context.

Improving the SE relies on improving the SNR. This can be accomplished by implementing communication schemes that more closely approach the underlying assumptions of Shannon's capacity formula (such as (i) Gaussian-shaped

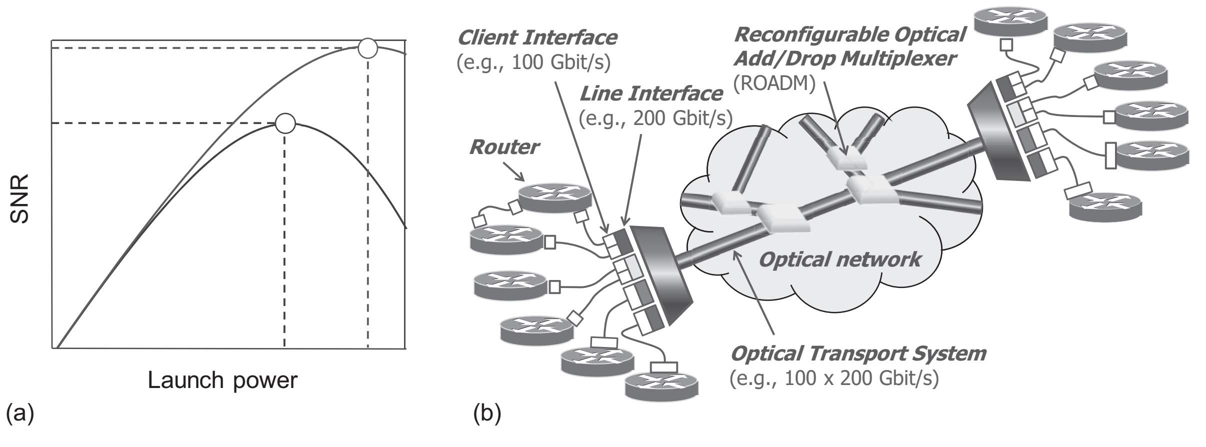


FIGURE 30.12 (a) Fibre non-linearity leads to an optimal signal launch power representing the balance between optical amplifier noise and NLIN. Techniques to mitigate NLIN aim at allowing for larger signal launch powers which results in a higher effective peak SNR. (b) General view of an optical transport mesh network.

modulation, approached by PCS, and (ii) close-to-ideal *forward error-correction coding, FEC* [51]), as well as by (iii) *increasing the effective SNR* of the received and fully electronically processed signal prior to entering the FEC decoder. Measures to improve the effective SNR in the context of fibre-optic communication systems include:

- lower-loss optical fibre
- lower noise amplification
- lower-non-linearity fibre (such that higher signal launch powers can be used to increase the SNR)
- non-linearity compensation techniques (digital and analog)
- improved DSP algorithms
- improved high-speed hardware (ADCs, DACs, modulators, drivers and pre-amplifiers, lasers, etc.)

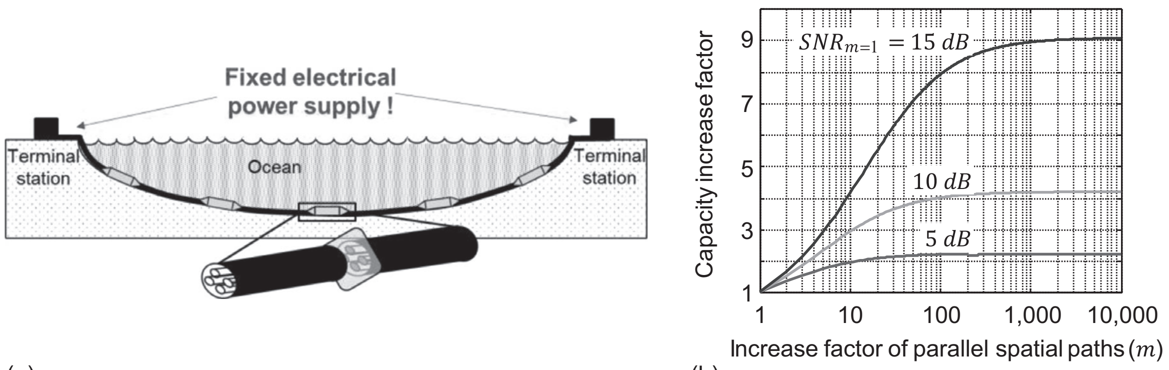
Importantly, all strategies to improve the SNR work *within the logarithm* of equation (30.1), and hence only show *logarithmic (sub-linear) gains* in the fairly high-SNR regimes typical of fibre-optic communication systems. This lets SNR improvements yield only very limited capacity gains. For example, a heroic SNR gain of 10 dB in a long-haul system operating at an SNR of 10 dB (SE_2 = 7 bits / s / Hz) will not even *double* system capacity. At a typical traffic growth rate of 60% per year, this buys the operator a mere 1.5 years. (Even at a very conservative traffic growth rate of 25% per year, a doubling in capacity is seen in only 3 years.) Note in this context that all of the above-mentioned techniques to improve the SNR try to eke out between a few tenths of a dB to a few dB in practice, far off any prospect of a 10-dB SNR improvement.

With reference to Figure 30.2a, virtually all metro and long-haul systems operate in the C-band, except for countries with a large installed base of dispersion-shifted fibre (DSF), such as Japan, which operate systems in the L-band to mitigate the more severe non-linear signal distortions on this fibre type [2]. Increasingly, systems using both C and L bands for a total of 9.6 THz of system bandwidth are being deployed (on fibre other than DSF) in order to double capacity relative to a single-band system. As per equation (30.1), capacity scales linearly with B, at least to first order when neglecting non-linear effects in massive WDM systems and engineering aspects such as power equalization across the band, or C/L band-splitters that are required because practically feasible optical amplifiers typically only cover a single amplification band. As such, increasing B is much more a question of amplifier bandwidths than of fibre bandwidths [2]. While the use of a third amplification band (such as the S-band) seems to be a commercial option, it is highly unlikely that commercial metro and long-haul transmission systems will increase bandwidths beyond the combined C+L+S-band. As a consequence, per-fibre WDM capacities are expected to saturate at or below 100 Tb s^{-1}, cf. Figure 30.8, which will force the deployment of parallel systems on M parallel spatial paths, e.g. fibre bundles, multi-core fibre, or even few-mode fibre (*'space-division multiplexing', SDM*) in the near future [2,6].

30.3.5 Sub-marine Transmission ($\gtrsim$ 5000 km)

Sub-marine systems [52] are limited by the amount of electrical supply power that can be fed into a sub-sea cable from both shores to power the optical amplifiers within the cable, as shown in Figure 30.13a [53]. Up until recently, cable supply powers were sufficient to operate typically up to 8 fibre pairs per cable at the non-linearity-optimized power, in complete analogy to terrestrial fibre-optic systems, cf. Figure 30.12a. This let sub-marine system designs essentially follow the design of terrestrial networks, with the difference that sub-marine lines can be designed in a much more controlled manner than terrestrial lines, e.g. using identical-length spans of identical-type fibre with the corresponding matched optical amplifiers in terms of gain and output power. Consequently, the capacity of a sub-marine cable (containing multiple parallel strands of fibre) has historically closely matched the capacity of a terrestrial long-haul WDM system, cf. Figure 30.8.

As WDM capacities per fibre saturate, sub-marine cables must deploy more parallel fibre strands to further increase cable capacities. Doing this in the usual way by retaining non-linearity-optimized launch powers quickly results in supply power limitations, though. A substantial increase in sub-marine cable capacity therefore asks for a different



(a) (b)

FIGURE 30.13 (a) A sub-marine transmission system is limited by the electrical supply power that can be fed into the cable from both shores to power the optical amplifiers within the cable. (b) First-order estimate of the capacity gain due to massive spatial parallelism for a fixed overall power constraint.

strategy, which follows directly from equation (30.1): By keeping the signal power *P* constant and increasing the number of parallel spatial paths *M*, the SNR per spatial path is reduced (and with it the fibre capacity per spatial path), but the linear pre-log factor of *M* over-compensates for the logarithmically lost per-fibre capacity. Figure 30.13b shows how the aggregate capacity of a system is increased when using *m* times the number of spatial paths of a reference system operating at the effective SNR parameterizing the curves and keeping fixed the overall available power, which is hence diluted among *m* times more parallel spatial paths. At a typical SNR of today's sub-marine cables of around 10 dB, a 4x capacity gain is predicted this way. A more accurate analysis [54] reveals even higher capacity gains when also optimizing span lengths, taking into account the reduced effects of NLIN with increasing power dilution among spatial paths, and considering reduced system bandwidths *B*, which is a power-saving option as broadband systems require passive gain flattening filters to equalize optical amplifier gain profiles and are hence not power-efficient. Future submarine cables are therefore expected to use relatively narrow bandwidths (such as half the C-band) and many parallel strands of fibre, or even advanced fibre such as MCF.

REFERENCES

1. J. Hecht, *City of Light: The Story of Fiber Optics*, Sloan Technology (2004).
2. P. J. Winzer, D. T. Neilson, and A. R. Chraplyvy, "Fiber-optic transmission and networking: the previous 20 and the next 20 years," *Opt. Express* 26, 24190–24239 (2018).
3. K. C. Kao and G. A. Hockham, "Dielectric-fibre surface waveguides for optical frequencies," *Proc. IEE* 113(7), 1151–1158 (1966).
4. Y. Tamura, H. Sakuma, K. Morita, M. Suzuki, Y. Yamamoto, K. Shimada, Y. Honma, K. Sohma, T. Fujii, and T. Hasegawa, "The first 0.14-dB/km loss optical fiber and its impact on submarine transmission," *J. Lightwave Technol.* 36(1), 44–49 (2018).
5. P. J. Winzer, "Would scaling to extreme ultraviolet or soft X-ray communications resolve the capacity crunch?," *J. Lightwave Technol.* 36, 5786–5793 (2018).
6. P. J. Winzer, "Spatial multiplexing in fiber optics: the 10x scaling of metro/core capacities," *Bell Labs Tech. J.* 19, 22–30 (2014).
7. P. J. Winzer, "High-spectral-efficiency optical modulation formats," *J. Lightwave Technol.* 20, 3824–3835 (2012).
8. C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.* 27, 379–423, 623–656 (1948).
9. J. P. Gordon, "Quantum effects in communications systems," *Proc. IRE* 50(8), 1898–1908, Sep. (1962).
10. S. Holevo, "Bounds for the quantity of information transmitted by a quantum communication channel," *Probl. Peredachi Inf.*, 9(3), 3–11 (1973).
11. R. S. Tucker, "Green optical communications—Part I: energy limitations in transport," *IEEE J. Sel. Topics Quantum Electron.* 17(2), 245–260 (2011).
12. P. J. Winzer and R.-J. Essiambre, "Advanced optical modulation formats," *Proc. IEEE* 94(5), 952–985 (2006).
13. A. Färbert, S. Langenbach, N. Stojanovic, C. Dorschky, T. Kupfer, and C. Schulien, "Performance of a 10.7-Gb/s receiver with digital equalizer using maximum likelihood sequence estimation," *Proc. European Conf. on Optical Comm. (ECOC)*, Th4.1.5 (2004).
14. D. McGhan, C. Laperle, A. Savchenko, C. Li, G. Mak, and M. O'Sullivan, "5120 km RZ-DPSK transmission over G652 fiber at 10 Gb/s with no optical dispersion compensation," *Proc. Optical Fiber Comm. Conf. (OFC)*, PDP27 (2005).
15. H. Sun, K-T. Wu and K. Roberts, "Real-time measurements of a 40 Gb/s coherent system," *Opt. Express* 16(2), 873–879 (2008).
16. X. Zhou, H. Liu, and R. Urata, "Datacenter optics: requirements, technologies, and trends," *Chin. Opt. Lett.* 15, 120008 (2017).
17. Concise overviews of various Ethernet speeds can be found on-line at: https://en.wikipedia.org/wiki/Ethernet_physical_layer; https://en.wikipedia.org/wiki/Terabit_Ethernet; https://en.wikipedia.org/wiki/100_Gigabit_Ethernet.
18. J. Goergen et al., "100Gb/s Electrical Signaling," IEEE 802.3 New Ethernet Applications Adhoc, March 2017; on-line: http://www.ieee802.org/3/ad_hoc/ngrates/public/17_03/goergen_nea_01a_0317.pdf.
19. K. Benyahya et al., "Multiterabit transmission over OM2 multimode fiber with wavelength and mode group multiplexing and direct detection," *J. Lightwave Technol.* 36(2), 355–360 (2018).
20. https://ethernetalliance.org/the-2019-ethernet-roadmap/.
21. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley (2018).
22. G. Agrawal, *Fiber-Optic Communication Systems*, Wiley (2010).
23. A. de la Oliva, J. A. Hernández, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Comm. Mag.* 54(2), 152-159 (February 2016).
24. A. Pizzinat, P. Chanclou, F. Saliou, and T. Diallo, "Things you should know about fronthaul," *J. Lightwave Technol.* 33, 1077–1083 (2015).
25. F. Effenberger, D. Cleary, O. Haran, G. Kramer, R. D. Li, M. Oron, and T. Pfeiffer, "An introduction to PON technologies," *IEEE Comm. Mag.* 45(3), S17–S25 (March 2007).
26. F. J. Effenberger, "The future of higher-speed PONs," *Proc. Asia Communications and Photonics Conference (ACP)*, 2018.
27. H. Rohde, E. Gottwald, A. Teixeira, J. D. Reis, A. Shahpari, K. Pulverer, and J. S. Wey, "Coherent ultradense WDM technology for next generation optical metro and access networks," *J. Lightwave Technol.* 32, 2041–2052 (2014).
28. N. Eiselt, J. Wei, H. Griesser, A. Dochhan, M. Eiselt, J.-P. Elbers, J. J. Vegas Olmos, and I. Tafur Monroy, "First real-time 400G PAM-4 demonstration for inter-datacenter transmission over 100 km of SSMF at 1550 nm," *Proc. Optical Fiber Comm. Conf.*, W1K.5D (2016).
29. A. Mecozzi et al., "Kramers–Kronig coherent receiver," *Optica* 3, 1218–1227, 2016.
30. X. Chen et al., "Kramers–Kronig receivers for 100-km datacenter interconnects," *J. Lightwave Technol.* 36, 79–89 (2018).

31. M. Chagnon, "Optical communications for short reach," *J. Lightwave Technol.* 37(8), 1779–1797 (2019).
32. C. R. Doerr and L. Chen, "Silicon photonics in optical coherent systems," *Proc. IEEE* 106(12), 2291–2301 (2018).
33. P. J. Winzer, "Transmission system capacity scaling through space-division-multiplexing: a techno-economic perspective," in *Optical Fiber Telecommunications VII*, Elsevier (2019).
34. G. Böcherer et al., "On joint design of probabilistic shaping and forward error correction for optical systems," *J. Lightwave Technol.* (2019).
35. J. Cho and P. Winzer, "Probabilistic constellation shaping for optical fiber communications," *J. Lightwave Technol.* 37(6), 1590–1607 (2019).
36. T. Okoshi and K. Kikuchi, *Coherent Optical Fiber Communications*, Springer (1988).
37. K. Kikuchi, "Fundamentals of coherent optical fiber communications," *J. Lightwave Technol.* 34(1), 157–179 (2016).
38. S. J. Savory, "Digital coherent optical receivers: algorithms and subsystems," *IEEE J. Sel. Top. Quantum Electron.* 16(5), 1164–1179 (2010).
39. S. L. I. Olsson, J. Cho, S. Chandrasekhar, X. Chen, E. C. Burrows, and P. J. Winzer, "Record-high 17.3-bit/s/Hz spectral efficiency transmission over 50 km using probabilistically shaped PDM 4096-QAM," *Proc. Optical Fiber Comm. Conf. (OFC)*, Th4C.5 (2018).
40. S. Chandrasekhar, B. Li, J. Cho, X. Chen, E. C. Burrows, G. Raybon, and P. J. Winzer, "High-spectral-efficiency transmission of PDM 256-QAM with parallel probabilistic shaping at record rate-reach trade-offs," *Proc. European Conf. on Optical Comm. (ECOC)*, Th.3.C.1 (2016).
41. J. Cho, X. Chen, S. Chandrasekhar, G. Raybon, R. Dar, L. Schmalen, E. Burrows, A. Adamiecki, S. Corteselli, Y. Pan, D. Correa, B. McKay, S. Zsigmond, P. Winzer, and S. Grubb, "Trans-Atlantic field trial using high spectral efficiency probabilistically shaped 64-QAM and single-carrier real-time 250-Gb/s 16-QAM," *J. Lightwave Technol.* 36(1), 103–113 (2018).
42. R.-J. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, and B. Goebel, "Capacity limits of optical fiber networks," *J. Lightwave Technol.* 28, 662–701 (2010).
43. R. Dar, M. Shtaif, and M. Feder, "New bounds on the capacity of the nonlinear fiber-optic channel," *Opt. Lett.* 39, 398–401 (2014).
44. J. Bromage, "Raman amplification for fiber communications systems," *J. Lightwave Technol.* 22, 79 (2004)
45. G. Agrawal, *Nonlinear Fiber Optics*, Elsevier (2013).
46. P. Poggiolini, G. Bosco, A. Carena, V. Curri, Y. Jiang, and F. Forghieri, "The GN model of fiber non-linear propagation and its applications," *J. Lightwave Technol.*, 32(4), 694–721 (2014).
47. R. Dar, M. Feder, A. Mecozzi, and M. Shtaif, "Properties of nonlinear noise in long, dispersion-uncompensated fiber links," *Opt. Express* 21, 25685–25699 (2013).
48. R. Dar and P. J. Winzer, "Nonlinear interference mitigation: methods and potential gain," *J. Lightwave Technol.* 35, 903–930 (2017).
49. Sorin Tibuleac and Mark Filer, "Transmission impairments in DWDM networks with reconfigurable optical add-drop multiplexers," *J. Lightwave Technol.* 28, 557–568 (2010).
50. S. Gringeri, B. Basch, V. Shukla, R. Egorov, and T. J. Xia, "Flexible architectures for optical transport nodes and networks," *IEEE Comm. Mag.* 48(7), 40–50 (July, 2010).
51. A. Leven and L. Schmalen, "Status and recent advances on forward error correction technologies for lightwave systems," *J. Lightwave Technol.* 32, 2735–2750 (2014).
52. A. Pilipetskii, "High capacity submarine transmission systems," in *Proc. Opt. Commun. Fiber Conf.*, 2015, Paper W3G. p. 5.
53. K. Takehira, "Submarine system powering," in *Undersea Fiber Communication Systems*, 2nd ed. J. Chesnoy Ed. New York: Academic (2016).
54. R. Dar, P. J. Winzer, A. R. Chraplyvy, S. Zsigmond, K.-Y. Huang, H. Fevrier, and S. Grubb, "Cost-optimized submarine cables using massive spatial parallelism," *J. Lightwave Technol.* 36(18), 3855–3865 (2018).

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

31

Visible Light Communications and LiFi

Cheng Chen, Mohamed Sufyan Islim, and Harald Haas

CONTENTS

31.1 Introductions 444
31.2 Communication Front-End Characterization 444
  31.2.1 Transmitters 444
    31.2.1.1 White-Coloured LEDs 445
    31.2.1.2 White-Coloured LDs 445
    31.2.1.3 Micro-Sized LEDs 445
    31.2.1.4 Resonant Cavity LEDs 446
    31.2.1.5 Organic LEDs 446
  31.2.2 Receivers 446
    31.2.2.1 PIN-PDs and APDs 446
    31.2.2.2 Single-Photon APDs 447
    31.2.2.3 PV Solar Cells 447
31.3 Indoor VLC Channel Characterizations 447
  31.3.1 Line-of-Sight Channel and Non-line-of-Sight Channel 447
  31.3.2 Deterministic Method 448
  31.3.3 Monte Carlo Method 448
  31.3.4 Frequency Domain Calculation 448
  31.3.5 Simplified Sphere Model 448
  31.3.6 Efficient Analytical Method 449
  31.3.7 Comparison of NLoS Channel Calculation Methods 450
31.4 Digital Modulation Techniques 450
  31.4.1 Single-Carrier Modulation Techniques 451
  31.4.2 OFDM-Based MCM 451
    31.4.2.1 DCO-OFDM 451
    31.4.2.2 Unipolar OFDM Techniques 452
    31.4.2.3 Hybrid OFDM Techniques 453
  31.4.3 Colour Modulation Techniques 453
31.5 Multiple-Input Multiple-Output Techniques in VLC and LiFi 454
  31.5.1 Spatial Multiplexing 454
  31.5.2 Repetition Coding 454
  31.5.3 Spatial Modulation 454
  31.5.4 Electrical-to-Optical Conversion Based on Multiple Light Sources 455
  31.5.5 MIMO Optical Source and Detector Deployment in VLC and LiFi 455
  31.5.6 Adaptation of MIMO Techniques 456
31.6 Co-channel Interference and Interference Mitigation Techniques in Multi-Cell LiFi Networks 456
  31.6.1 Access Point Layout and SINR Characterizations 456
  31.6.2 Interference Mitigation Techniques 457
31.7 Conclusion 458
References 458

31.1 Introductions

With the increasing number of ‘smart’ devices and the emergence of the internet-of-things (IoT), a significant increase of wireless data traffic is expected in the near future [1]. Given the overcrowding which already exists in the radio frequency (RF) spectrum, it will become increasingly difficult to fulfil the future mobile data traffic demand. Exploring new electromagnetic spectra for wireless transmission has led to the research and development of visible light communication (VLC). Based on this technology, a new type of wireless access network has been developed, which is known as LiFi (Light-fidelity) [2]. A LiFi network uses multiple indoor luminaries with signal processing capabilities as wireless access points (APs), which are able to communicate with each other and jointly serve nearby mobile devices. This forms a bidirectional, wide coverage and multi-functional wireless system [3]. Compared to conventional RF access networks, LiFi networks are more secure and have superior performance in a dense user scenario [4,5]. In this chapter, a series of important technologies that can be used in VLC and LiFi networks are briefly introduced.

31.2 Communication Front-End Characterization

Electrical-to-optical conversion is performed by the transmitter (Tx) of a VLC system. Solid-state lighting sources such as light emitting diodes (LEDs) and laser diodes (LDs) are possible choices for VLC. Photodetectors used in optical wired and wireless communications such as silicon based positive-intrinsic-negative photodiodes (PIN-PDs) and avalanche PDs (APDs) are suitable candidates for VLC due to their relatively high responsivity between wavelength $\lambda = 320\,\text{nm}$ and $\lambda = 1100\,\text{nm}$. This section characterizes front-end devices in VLC.

31.2.1 Transmitters

The Tx end of a VLC system is expected to meet the illumination requirements. A VLC Tx can be characterized by its radiation pattern, current-voltage-luminance characteristics, frequency response profile and spectral response. LEDs are prominent candidates for VLC due to their low cost and vast deployments as illumination sources. LEDs have started to replace other illumination sources such as incandescent and fluorescent light sources [6]. This is mainly due to their high luminance efficacy and their significant energy consumption reduction compared to incandescent and fluorescent light sources. The radiation pattern of LEDs is widely approximated by a generalized Lambertian model. The radiation pattern is characterized by the semi-angle at half power representing the field of view of a LED, $\phi_{1/2}$. Commercially available LEDs have a Lambertian order around $m = 1$, which corresponds to a semi-angle $\phi_{1/2} = 60°$. The radiation pattern of a LED can be modified using optical components such as reflectors, lenses and diffusers. An LED-based luminaire uses multiple LEDs to satisfy the lighting requirements. A desired radiation pattern can be achieved using multiple LEDs and optical components to minimize interference and increase the system throughput based on the desired application. This allows for an improved communication coverage and an increase in the spectral efficiency (SE) per unit area. LDs are much more directive and their radiation pattern is generally modelled with a Gaussian distribution.

The current-voltage (I-V) and luminance-current (L-I) transfer functions of a light source underpins the performance of the VLC system. An ideal light source is assumed to have linear I-V and L-I transfer functions with double side-clipping points representing the turn-on voltage and the saturated optical power. However, the I-V and L-I transfer functions of commercially available devices are mostly non-linear. The non-linearity of the Tx restricts the dynamic region where the transmitted signal can swing. The non-linearity of the I-V and L-I characteristics distorts the modulation signal and impose degradations on the signal-to-noise ratio (SNR). The L-V characteristic function of a typical light source is shown in Figure 31.1. An input signal with a Gaussian distribution is demonstrated to result in a truncated Gaussian distribution output due to the nonlinearity of the light source. The transmitted signal power is the average power of the Tx within the linear region of the dynamic range. The non-linearity of LEDs has been empirically studied in various studies [7–10]. A complete modelling of memory-less non-linear distortion based on orthogonal frequency division multiplexing (OFDM) in intensity modulation and direct detection (IM/DD) systems is presented in [11]. Signal pre-distortion was considered to mitigate the effects of non-linearity. However, the I-V and L-I transfer functions of a device are not static and vary with the frequency of the transmitted signal [12]. Furthermore, the I-V and L-I characteristics are affected by the temperature variations of the device. Therefore, thermal management of VLC Txs is essential to mitigate any performance degradation and ensure the reliability of the system.

The frequency response of the Tx is directly proportional to the achievable data rate. The 3-dB bandwidth is commonly used to contrast the frequency profile of a device. However, the roll-off of the frequency response is important since the modulation bandwidth of VLC Txs can exceed the 3-dB bandwidth

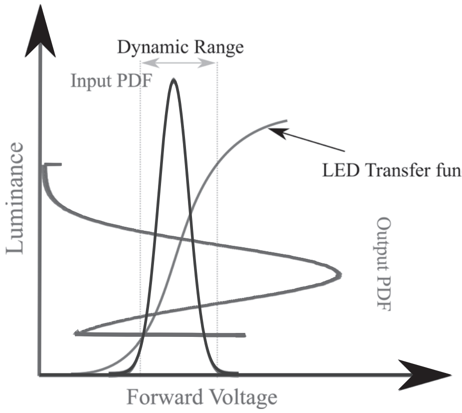


FIGURE 31.1 The voltage-luminance characteristic of semiconductor-based light source. Input signal with a Gaussian probability density function (PDF) and an output signal with a truncated Gaussian PDF are considered.

when equalization techniques are used. Therefore, other metrics such as the 6-dB bandwidth are also used to reflect on the frequency response roll-off of a device. The frequency response of an LED is commonly approximated with a low pass filter. The 3-dB bandwidth of commercially available LEDs ranges between a few kHz and a few MHz, whereas the modulation bandwidth of LDs ranges between a few MHz and a few GHz. The frequency response of LEDs is mostly affected by the carrier lifetime and the parasitic capacitance of the LED junction. A relative comparison of the characterizations of PC-LEDs, LDs, m-LEDs and RC-LEDs based on bandwidth, optical power, spectral width, divergence angle and the achieved peak data rate is presented in Figure 31.2. The spectral width is given by the full-width at half-maximum (FWHM). Devices with small spectral width have the potential to increase overall throughput in a wavelength division multiplexing (WDM) system. On one hand, devices with higher bandwidth are desired as the modulation bandwidth is directly proportional to the achievable data rate. This shows the potential for laser-based Txs in VLC systems. On the other hand, devices with high optical power are desired to relax the alignment constraints and enable longer transmission distances. Divergence angle can be modified using optical components.

31.2.1.1 White-Coloured LEDs

Monochromatic LEDs have a broad spectral width with the FWHM of commercially available LEDs ranging from 25 to 100 nm. Multiple streams can be transmitted using multiple LEDs emitting at different wavelength bands which increases the throughput in a WDM system. White illumination can be achieved using a combination of red, green and blue LEDs (RGB-LEDs). The combined spectral response of multi-colour LEDs can result in white illumination based on the colour-mixing concept that dates back to James Clarke Maxwell in 1861. Optical filters with spectral responses matching the wavelength bands of the LEDs are required in WDM systems. This is essential to reduce the interference resulting from the overlap of the adjacent wavelength bands of the used LEDs.

Yellow phosphor pumped with a blue-coloured LED is most commonly used to generate white illumination due to the low cost of manufacturing. The process of converting high-energy short wavelengths into longer wavelengths is known as down-conversion. The white light output of a phosphor-coated LED (PC-LED) has two components: a relatively narrow band in the blue spectrum (ranging from wavelength $\lambda = 425$ to $\lambda = 450$) and a relatively broad band in the in the spectrum ranging from $\lambda = 500$ to $\lambda = 650$ [13]. The yellow component endures a slow response which results in reducing the 3-dB bandwidth of the bare blue LED. A blue bandpass filter is commonly used at the receiver (Rx) to filter out the slow response of the phosphor. The 3-dB bandwidth of commercially available PC-LEDs is around a few hundreds of kHz. However, this value can be increased to a few MHz when a blue filter is used [14]. Recent results reported 2.32 Gbps using PC-LED with the aid of blue filters and two-staged linear software equalizers [15]. The long photoluminescence response of the phosphor has motivated the pursuit of fast down-conversion materials which allow for a better white light quality and a higher 3-dB bandwidth such as quantum dots [16,17].

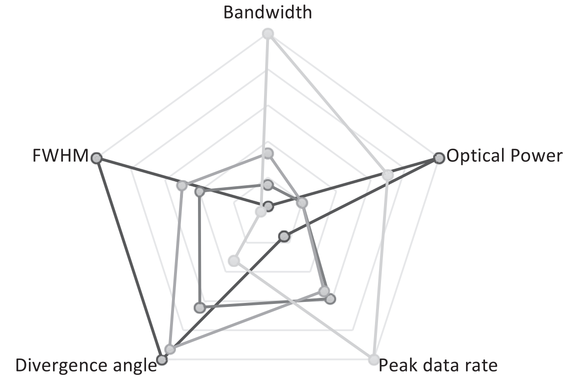


FIGURE 31.2 A relative comparison of the characterizations of PC-LEDs, LDs, m-LEDs and RC-LEDs.

31.2.1.2 White-Coloured LDs

There has been an increasing interest in using LDs for general lighting due to their optical power efficiency at high current densities [18]. LDs have also gained an increasing interest in VLC recently, mainly due to their superior frequency response compared to LEDs reaching a 6.8 GHz of 3-dB bandwidth [19]. Recent results reported 15 Gbps using a blue LD [20] with a 2.5 GHz modulation bandwidth. In addition, LDs can have a spectral width as narrow as 1 nm [21]. This allows LDs to be used for a higher dense deployment of the available visible spectrum. It has been reported that 100 Gbps can be achieved using state-of-the-art sources [22]. The generation of white light using LDs is not straightforward due to eye safety concerns and speckle issues. The coherence of the LDs needs to be destroyed before an LD can be considered for illumination. Multiple LDs emitting at red, green and blue wavelengths can be used to generate white illumination [22]. Recent results reported more than 20 Gbps using RGB LDs [23]. Remote phosphor can be excited with a blue laser to achieve white illumination in a similar fashion to PC-LEDs achieving a data rate around 5.2 Gbps [24]. A combination of a conventional red phosphor and Perovskite nanocrystals were considered as down-converting materials for blue LDs, achieving a peak data rate around 2 Gbps with a colour rendering index (CRI) of 89 and a correlated colour temperature (CCT) around 3200 K [25]. A violet LD is used with red, green and blue phosphor to achieve white illumination at a CRI of 90, and CCT around 2700 K [26] achieving a data rate above 1 Gbps. Figure 31.2 demonstrates the potential of LDs in future VLC systems compared with other light sources.

31.2.1.3 Micro-Sized LEDs

Micro-sized GaN LEDs (m-LEDs) attracted an increasing interest in the last few years due to their high 3-dB bandwidth exceeding 800 MHz [27]. Recent results reported a data rate of 7.91 Gbps using a violet m-LED with an active area of 435 μm^2 and a 3-dB bandwidth of 655 MHz [12]. A reduction in junction size results in a high current density which allows for a short carrier lifetime. However, the decrease in LED sizes results in a reduction in the optical power of m-LEDs. Multiple m-LEDs can be used to increase the optical power. However, thermal management is essential to avoid nonlinearity and to avoid the degradation in the m-LED lifetime.

31.2.1.4 Resonant Cavity LEDs

Resonant cavity LEDs (RC-LEDs) have a higher modulation bandwidth compared with conventional LEDs. A resonant cavity at the active region of RC-LED allows for an increased radiant intensity and improved directionality. Recent results reported a data rate of 8.76 Gbps using an improved equalizer with a RC-LED that has a 100 MHz 3-dB bandwidth [28].

31.2.1.5 Organic LEDs

Organic LEDs (OLEDs) are commonly used in display systems due to their low manufacturing cost, flexibility and wide radiation pattern. However, a major drawback of using OLEDs for general lighting is their lifetime and brightness. Recently, there has been an interest in evaluating and characterizing OLEDs for VLC [29]. The 3-dB bandwidth of OLEDs is in the order of a few hundred kHz due to their low carrier mobility. However, the 3-dB bandwidth of OLEDs can be improved by optimizing the used organic material [30].

31.2.2 Receivers

Photodetectors are characterized by the light collection pattern, frequency response, spectral response, detection area, sensitivity and noise figure. Commercially available PIN-PDs and APD-PDs are the most commonly used Rxs for VLC. Single APDs (SPADs) have recently attracted more interest due to their high sensitivity [31,32]. Detectors capable of optical to electrical conversion such as photovoltaic (PV) cells have been recently considered for VLC systems [33]. Image sensors such as charge-coupled devices and complementary metal oxide semiconductor (CMOS) cameras have been used for optical camera communications (OCCs). However, the achievable data rates in OCCs are limited to a few hundreds of kbps due to the low frames-per-second rate of these image sensors. This section is focused on photodiodes and PV cells as detectors due to their capabilities in achieving multi-Gbps VLC.

31.2.2.1 PIN-PDs and APDs

The light collection pattern of a PD is similar to the radiation pattern of LEDs which can be approximated with a generalized Lambertian model. The light collection pattern of the PD can be modified using optical elements that can change the field-of-view (FoV) of the PD based on the desired scenario. For example, it would be suitable to decrease the FoV in an interference-limited VLC system, whereas it would be more suitable to increase the FoV in a low-optical-power-limited VLC system. Optical elements such as concentrator lenses can be used to improve the light collection. The gain of the concentrator can be given as follows [21]:

$$g(\psi) = \begin{cases} \dfrac{n^2}{\sin^2 \Psi_c}, & 0 \le \psi \le \Psi_c \\ 0, & \psi > \Psi_c, \end{cases} \tag{31.1}$$

where ψ is the angle of incidence; n is the refractive index of the concentrator; and Ψ_c is the FoV of the concentrator. The light collection pattern can also be modified using multiple detectors which can be placed on a non-planar base with at least a single PD at each facet to allow for an angle diversity photoreceiver [34,35]. This allows the photoreceiver to suppress interfering signals and background light noise. Low-power CMOS-based application-specific integrated circuit (ASIC) PDs were designed for LiFi applications [2]. The ASIC chip has an array of 49 APDs with versatile spatial and spectral properties. Signal combing schemes can be used to maximize the SNR of such photoreceivers at the cost of increased implementation and operation complexity.

Spectral bandwidth with an FWHM as narrow as 1 nm is commercially available with a transmittance above 80% [21]. Photodetector with high sensitivity are desired at the wavelength of interest. Silicon PDs are the most commonly used PDs due to their spectral responsivity in the visible spectrum which peaks at wavelengths longer than the red wavelength. A high reverse voltage is applied on PDs to achieve a multiplicative gain. This allows for an avalanche of photocurrent to be generated when triggered by a number of photons. Therefore, an APD is essentially a PD that has an internal gain. The gain results in a higher sensitivity for APDs compared with PIN-PDs. The high sensitivity of APDs promotes them to applications where low optical power is expected at the detector such as long range VLC systems and underwater optical wireless communications. Thermal noise and shot noise dominate the variance of the noise at the PD. Shot noise is affected by both the background intensity and the optical source intensity. Optical and electrical filters are commonly used at the Rx side to mitigate out of band noise [36]. The shot noise increases with the internal gain due to the multiplicative process in APDs. This limits their performance when strong background light is present at the detector. Optical power budgeting can be achieved using lenses, optical bandpass filters and neutral density filters. In addition, automatic gain controllers can be used to adapt the multiplicative gain of the APD based on the achieved SNR.

The frequency response of a photodetector is inversely proportional to the active detection area. Increasing the detection area results in an increase in the received optical power. However, this increases the Rx noise and decreases the 3-dB bandwidth of the detector due to the increased capacitance. Most of the recent research efforts on front-ends in VLC focused on the Tx. This is due to the fact that, in most cases, the main limiting factor of high-speed VLC systems is the light source. However, recent results on high-speed VLC Txs reported limitations in the modulation bandwidth of the system due to photodetectors [12,19]. It is interesting to note that photodetectors with higher modulation bandwidth could have been used in these studies. However, such PDs have a very small detection area. For example, the detection area of a commercially available 1 GHz PIN-PD is 0.283 mm^2 (ALPHALAS UPD-300-SP) is 7 times larger than the detection area of 10 GHz PIN-PD (ALPHALAS UPD-30-VSG-P). PDs with a small detection area limit the application of VLC to a very narrow FoV which requires tedious alignment and reduces the system reliability and practicality. There is an increasing demand to improve the 3-dB bandwidth of photodetectors while maintaining a relatively large detection area that would allow for VLC and LiFi applications to be feasible. Decoupling the light collection mechanism from the light detection has resulted in novel optical components that could improve the light detection area without significantly compromising the 3-dB bandwidth of the PDs [37,38].

31.2.2.2 Single-Photon APDs

Single-photon APD (SPAD) is a PD that is reverse biased beyond the breakdown voltage. SPADs are photon-counters that operate at the photon-level sensitivity. A single detected photon with a SPAD creates an avalanche of photocurrent [39]. SPADs have the highest sensitivity when compared with other types of PDs. Therefore, they are considered for scenarios where the received optical power is severely attenuated, such as downhaul monitoring [40]. The bandwidth of SPADs is limited by the dead time which quantifies the recovery time required for the SPAD to detect newly arriving photons. Another major drawback of SPADs is their non-linear response due to limitations in the maximum number of photons an SPAD can detect. A number of SPADs can be used in an array to mitigate this issue. Recent results reported 200 Mbps using an array of SPADs [41].

31.2.2.3 PV Solar Cells

PV solar cells have recently gained an increasing interest as VLC photodetectors [33, 42–44]. PV cells offer an interesting solution for backhauling and IoT applications as they enable simultaneous wireless information and power transfer [45]. A major drawback of using commercially available PV cells as photodetectors is their limited 3-dB bandwidths which are in the order of a few hundreds of kHz. However, recent results reported a data rate of 0.5 Gbps for a VLC system using a GaAs PV cell with a 3-dB bandwidth of 24.5 MHz [43]. In addition, organic PV cells with a 3-dB bandwidth of 1.3 MHz were reported to achieve a data rate of 34.2 Mbps [44] while simultaneously harvesting 0.43 mW power.

31.3 Indoor VLC Channel Characterizations

In a VLC or LiFi system, a light signal is transmitted through the indoor free-space channel. In this process, a fraction of the signal energy is captured by the optical detector, while the remaining signal energy is lost. This is known as path loss [13]. In addition, a reflective indoor environment creates a multi-path propagation, which leads to multiple signal energy streams reaching the detector at different times. This multipath effect will lead to different path losses at different frequencies, which is known as frequency-selectivity. An optical signal experiences diffused reflection when it hits rough indoor interior surfaces, which is not the same for low-frequency RF signals experiencing specular reflections. These facts make the channel models for RF systems not applicable in VLC/LiFi systems. Consequently, a series of unique approaches have been developed to characterise indoor VLC channels.

31.3.1 Line-of-Sight Channel and Non-line-of-Sight Channel

Indoor VLC channels can generally be divided into two categories: line-of-sight (LoS) channels and non-line-of-sight (NLoS) channels [46]. As shown in Figure 31.3, LoS channels account for the signal propagation through the path directly from the Tx to the Rx. Generally, it depends on the Tx/Rx separation, the angular intensity/orientation of the Tx source and the angular sensitivity/orientation of the Rx detector. With appropriate alignment, a robust LoS channel can be established to provide a reliable connection. The path loss of a LoS channel can be characterized by [47]:

$$H_{\mathrm{LoS}} = \frac{(m+1)A_{\mathrm{rx}}}{2\pi D^2}\cos^m\phi\cos\psi\mathbf{1}_{\mathrm{v}}(\phi,\psi), \qquad (31.2)$$

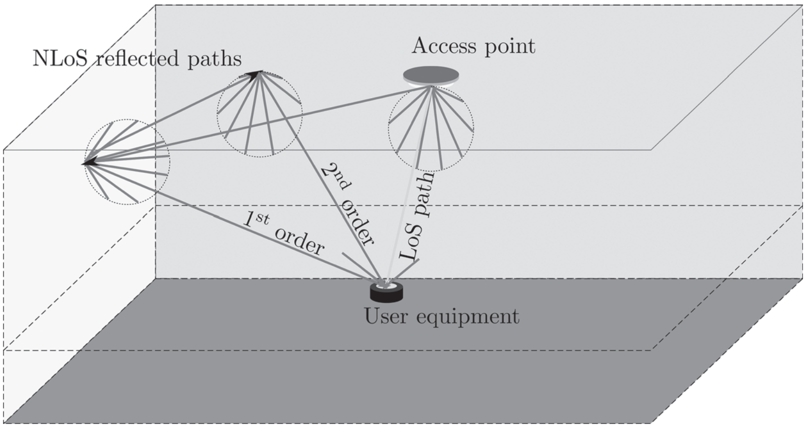


FIGURE 31.3 VLC signal propagation in an indoor environment.

where D represents the Tx/Rx separation, ϕ denotes the radiant angle at the Tx, ψ denotes the incident angle at the Rx, A_{rx} is the physical area of the Rx detector, m is the Lambertian mode number which is related to the half-power semi-angle of the optical source $\phi_{1/2}$ and $\mathbf{1}_{\mathrm{v}}$ is the visibility factor, which equals zero if ϕ exceeds $\pi/2$ or ψ exceeds the FoV of the Rx ψ_{FoV}.

In addition to the LoS channel, a fraction of the signal energy is incident to and then reflected by surrounding objects, as shown in Figure 31.3. A fraction of the reflected signals are captured by the Rx detector. The accumulation of detected signals from reflections forms the NLoS channel. The surfaces of most indoor objects such as walls and furniture are not smooth relative to the short-wavelength optical signals. Consequently, these surfaces generally lead to diffused reflections. Occasionally, some other objects such as windows and mirrors may cause specular reflections, but the probability that the specular reflected signal can reach the Rx is very low. The NLoS channel significantly depends on the characteristics of the indoor environment, which makes the calculation of the NLoS channel require high computational complexity. In the following sections, we introduce a number of useful methods to calculate NLoS channels in VLC and LiFi systems.

31.3.2 Deterministic Method

An intuitive way of calculating the NLoS channel is using ray-tracing techniques. An accurate deterministic ray-tracing simulation method has been proposed in [46]. In this method, the indoor surfaces that cause reflections are discretized and divided into a large number of small blocks for the convenience of numerical evaluation. Then, the power transfer interactions (path loss and delay) between blocks are calculated. Based on these calculated quantities, the NLoS channel impulse response corresponding to different order of reflections can be approximated by a finite sum of scaled delta functions, which can be evaluated using a recursive algorithm [46]. Here, the order of reflections means the number of reflections that the signal light undergoes. Note that the spatial discretization in the deterministic method also causes discretization in time. In order to retrieve the smooth piecewise-continuous impulse response waveform, time is divided into bins of small width and the power of the signals falling into the same time bin is accumulated. With further investigation, a more efficient iterative algorithm has been proposed [48]. In this improved method, the calculation results for the impulse response corresponding to the kth order reflections are reused to calculate the impulse response corresponding to the $k+1$th order reflection, which leads to a faster calculation of NLoS channel impulse response. The aforementioned method was designed for narrow line-width infrared wireless transmission. The authors of [49] have extended the method for signals with a broader spectrum in the visible light range.

The deterministic method is highly flexible, which means it is applicable to indoor environments with reflecting objects of any shape. However, the accuracy depends on the ratio of the block size and the temporal smoothing time bin width. The computation time scales significantly with the number of blocks. When the room interior surface area gets larger or the size of blocks gets smaller, the entire indoor surfaces are sliced into a greater number of blocks. Furthermore, the computation time also depends on the maximum order of reflection that is included. Consequently, when considering multiple NLoS channels with high bandwidth deployed in a large indoor environment, it will take several hours, or even days, to finish the calculation using the deterministic ray-tracing method.

31.3.3 Monte Carlo Method

In order to calculate the NLoS channel more efficiently, another Monte Carlo ray-tracing approach has been proposed [50]. In this method, the propagation of a light ray is traced. In this process, the emission direction of the light ray is randomly selected based on the radiation pattern of the source. When the light ray meets any reflective surface, it is attenuated and reflected in another random direction. At the same time, the power contribution to the Rx detector and the corresponding delay are calculated and recorded. The tracing ends when either the delay reaches the maximum tracing time or the light ray experiences the maximum order of reflections. Finally, if the tracing process is repeated with a large number of iterations, the accumulation of the power reaching the Rx detector eventually forms the impulse response of the NLoS channel.

Compared to the deterministic method, the Monte-Carlo based method is more efficient and the computational complexity is lower. However, the randomness in propagation direction introduces some ‘noise’ in the channel impulse response results. In order to minimize this effect, a large number of iterations are required to average out this ‘noise’. A characterization of the ‘noise’ has also been conducted in [51].

31.3.4 Frequency Domain Calculation

In most of the ray-tracing based approaches, the NLoS channel impulse responses are calculated in the time domain. In contrast, Henrik [52] has come up with an approach to calculate the NLoS channel transfer function in the frequency domain. The main advantage of this approach is that it converts most of the convolution operations in the time domain into matrix multiplication operations in the frequency domain, which is more convenient to implement. In addition, it has been shown that the computational complexity of this method is much lower compared to the time-domain method. This means that it requires less time to finish the calculation. Furthermore, this frequency-domain method can be easily extended to the case of calculating multiple NLoS channels with various Tx/Rx pairs, and it can include an unlimited number of reflections. However, the computation time scaling with the number of block elements still exists.

31.3.5 Simplified Sphere Model

The ray-tracing based method is known to offer a very accurate NLoS channel impulse response with a relatively high computational complexity. In some cases, the transmission channel is limited by other factors, such as front-end bandwidth.

In such conditions, a coarse estimation of the NLoS channel is sufficient. A low complexity method considering a simplified sphere physical model has been proposed to calculate the NLoS channel calculation [53]. In this method, the cubic indoor environment has been approximated by the interior of a spherical space with equivalent size and reflectivity. The main advantage of this method is that it ends up with a simple closed form expression. However, when the indoor environment becomes less similar to the interior of a spherical space, the accuracy of the simplified sphere model decreases significantly. In some cases, the magnitude of the NLoS channel calculated using the simplified sphere model can be 10 dB smaller than that simulated using the frequency domain method [52].

31.3.6 Efficient Analytical Method

In many studies on VLC or LiFi, the performance evaluation is conducted at system level. In this type of study, the randomness of user locations in a relatively large indoor space is taken into account. In order to evaluate the empirical statistical result from this randomness, a large number of system performance samples with various user locations are required, which may lead to the accurate calculation of a massive number of NLoS channels. Due to the significant increase in the number of channel calculations, it is normally impractical to use ray-tracing based methods to finish the computation within a reasonable time. On the other hand, the use of the simplified sphere model leads to inaccurate results. An analytical approach achieving a compromise between complexity and accuracy has been proposed in [54].

The general idea of this method is to decompose the NLoS channel into multiple components with simpler light propagation geometry, as shown in Figure 31.4a. This decomposition is based on the number of reflections and the surfaces that the optical signal undergoes. The impulse responses of a few dominant components can be efficiently calculated using tractable expressions. Finally, a reasonably accurate NLoS channel impulse response can be evaluated from the superposition of the component impulse responses within one second. The main advantage of this method is its low computational complexity relative to the ray-tracing based methods. At the same time, the effects of a large number of parameters are included in the calculation, which leads to better accuracy and flexibility relative to the simplified sphere model.

This analytical method is motivated by the ceiling-bounce model proposed in [55], which considers a single reflection channel in a special case. In [54], this single reflection channel analytical expression is extended to a general case, as shown in Figure 31.4b. The analytical expression of the impulse response of this single-reflection channel is concluded as:

$$\hbar^{\vec{o}_\mathrm{s},\,\vec{o}_\mathrm{r}}_{\{m,\,A_\mathrm{rx},\,\rho,\,L_\mathrm{s},\,L_\mathrm{r},\,L_\mathrm{b},\,\psi_\mathrm{FoV}\}}(t) = \frac{\rho(m+1)L_\mathrm{s}L_\mathrm{r}A_\mathrm{rx}}{2\pi^2} \times \mathcal{U}\left(t - \frac{L_0}{c}\right)\int_0^{2\pi} \mathfrak{r}_\theta(t)\,\mathfrak{f}\left(\mathfrak{r}_\theta(t),\,\theta\right)\frac{\mathrm{d}\mathfrak{r}_\theta(t)}{\mathrm{d}t}\mathrm{d}\theta, \tag{31.3}$$

where

$$\begin{aligned}\mathfrak{f}(r,\theta) = {} & \left(r\left(\tilde{x}_\mathrm{s}\cos\theta + \tilde{y}_\mathrm{s}\sin\theta\right) + \frac{L_\mathrm{s}L_\mathrm{b}\tilde{x}_\mathrm{s}}{L_\mathrm{s}+L_\mathrm{r}} + L_\mathrm{s}\tilde{z}_\mathrm{s}\right)^m \\ & \times\left(r\left(\tilde{x}_\mathrm{r}\cos\theta + \tilde{y}_\mathrm{r}\sin\theta\right) - \frac{L_\mathrm{r}L_\mathrm{b}\tilde{x}_\mathrm{r}}{L_\mathrm{s}+L_\mathrm{r}} + L_\mathrm{r}\tilde{z}_\mathrm{r}\right)\mathbf{1}_\mathrm{FoV} \\ & \times\left(r^2 + \frac{2L_\mathrm{s}L_\mathrm{b}r\cos\theta}{L_\mathrm{s}+L_\mathrm{r}} + L_\mathrm{s}^2 + \frac{L_\mathrm{s}^2L_\mathrm{b}^2}{\left(L_\mathrm{s}+L_\mathrm{r}\right)^2}\right)^{-\frac{m+3}{2}} \\ & \times\left(r^2 - \frac{2L_\mathrm{r}L_\mathrm{b}r\cos\theta}{L_\mathrm{s}+L_\mathrm{r}} + L_\mathrm{r}^2 + \frac{L_\mathrm{r}^2L_\mathrm{b}^2}{\left(L_\mathrm{s}+L_\mathrm{r}\right)^2}\right)^{-2},\end{aligned} \tag{31.4}$$

$$L_0 = \sqrt{L_\mathrm{b}^2 + \left(L_\mathrm{s}+L_\mathrm{r}\right)^2}, \tag{31.5}$$

$$\begin{aligned}\mathfrak{c}_\theta(t) = {} & 4L_\mathrm{s}L_\mathrm{r}\left(c^2t^2 - L_\mathrm{b}^2\cos^2\theta\right)\left(c^2t^2 - L_0^2\right) \\ & +\left(L_\mathrm{s}-L_\mathrm{r}\right)^2\left(c^2t^2 - L_0^2\right)^2.\end{aligned} \tag{31.6}$$

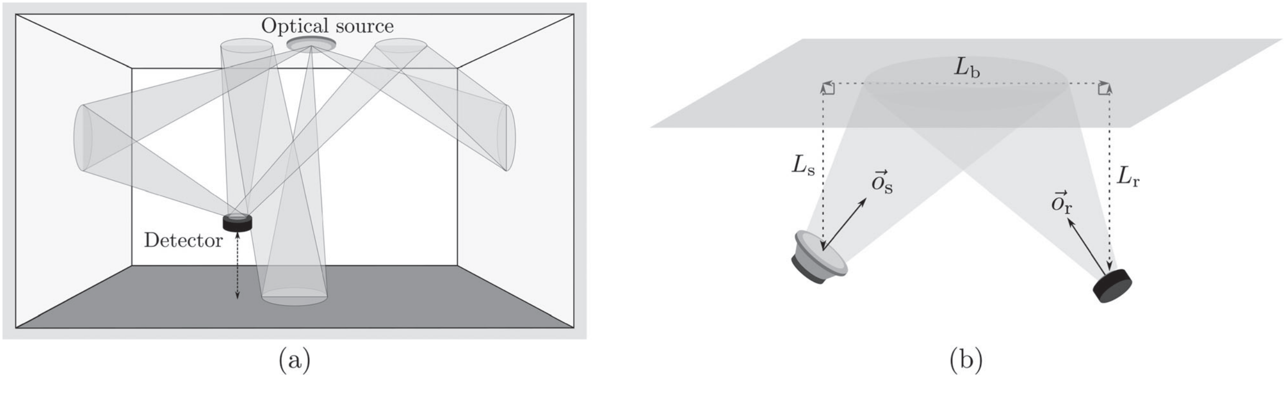
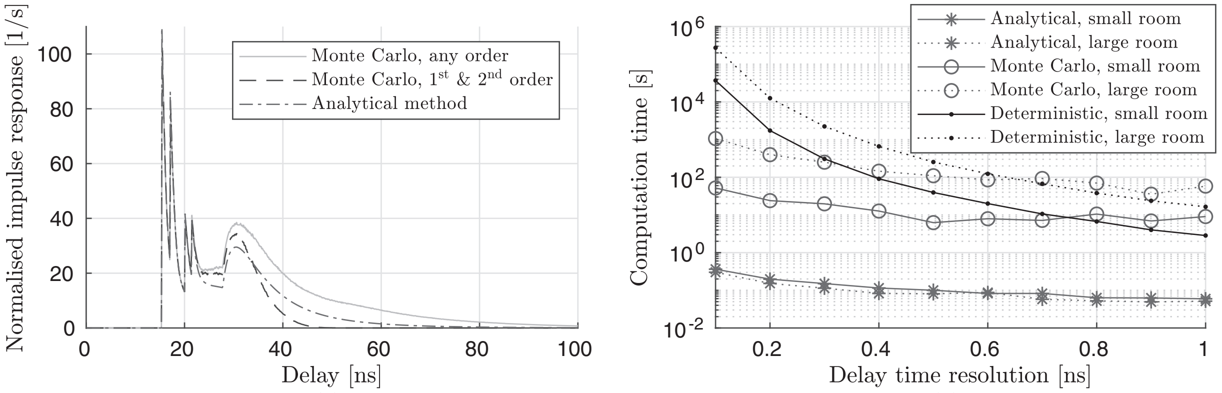


FIGURE 31.4 (a) Demonstration of NLoS channel decomposition. Red: Tx-to-wall-to-Rx (first order reflection); Tx-to-floor-to-ceiling-to-Rx (second order reflection); Tx-to-wall-to-ceiling-to-Rx (second order reflection). (b) Generalized single reflection NLoS channel geometry. L_b is the separation between the projections of Tx/Rx position on the reflecting surface plane. L_s (L_r) is the distance between the projection point and the corresponding position of Tx (Rx). $\vec{o}_\mathrm{s}$ and $\vec{o}_\mathrm{r}$ are orientation vectors of Tx source and Rx detector with a modulus of unity.



(a) (b)

FIGURE 31.5 (a) NLoS channel impulse response generated using analytical method. (b) The simulation time comparison between deterministic, Monte Carlo and analytical methods. The results are extrapolated from Figure 31.13 (a) and Figure 31.15 in Ref. [54].

In (31.3–31.6), ρ represents the reflectivity of the reflective surface and c represents the speed of light. The geometric parameters $\vec{o}_\mathrm{s}$, $\vec{o}_\mathrm{r}$, L_s, L_r and L_b are illustrated in Figure 31.5. In addition, the operator $\mathcal{U}(u)$ is the unit step function and $\mathbf{1}_{\mathrm{FoV}}$ is an indicator function, which guarantees that only the signals propagating within the FoV of Rx and light source coverage contribute to the CIR.

Based on, the expressions for the component impulse responses of various categories can be derived. Finally, the complete NLoS channel impulse response can be approximated by the superposition of the dominate component impulse responses. The accuracy and the computation time of the analytical method are demonstrated in Figure 31.5.

31.3.7 Comparison of NLoS Channel Calculation Methods

Based on the aforementioned introductions, the characteristics of various NLoS channel calculation methods have been concluded in Table 31.1. It can be concluded that when the computation involves a fewer number of NLoS channels and the calculation accuracy requirement is relatively high, the ray-tracing based methods (deterministic, Monte-Carlo and frequency domain calculation) are suitable candidates. If the computation involves a large number of NLoS channels and the accuracy requirement is medium or low, the simplified sphere model and analytical methods are more suitable options.

TABLE 31.1

Comparison of Various NLoS Channel Calculation Methods

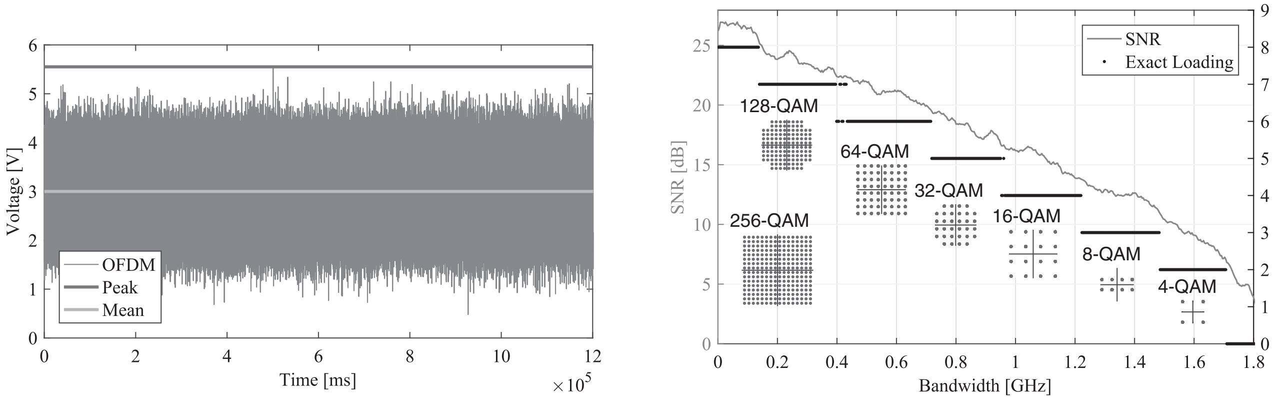
Calculation Method	Accuracy	Complexity	Flexibility
Deterministic	Very high	Very high	High
Monte-Carlo	High	Medium	High
Frequency domain	Very high	Medium	High
Simplified Sphere model	Low	Very low	Low
Analytical calculation	Medium	Low	Medium

31.4 Digital Modulation Techniques

Incoherent light sources are widely believed to be the prominent choice for VLC Txs due to their low cost and simple system design. This restricts the modulation techniques in VLC to IM/DD schemes. A spatial diversity can be achieved using IM/DD due to the relatively large-area detectors compared to the wavelength of the detected light. In comparison with coherent-based RF modulation techniques this mitigates the multipath fading and simplifies the VLC system design.

Single-carrier modulation (SCM) techniques such as on-off keying (OOK); pulse width modulation (PWM); pulse position modulation (PPM) and pulse amplitude modulation (PAM) can be easily implemented in VLC. The front-end frequency selectivity and the multipath effects of the visible light channel result in a high inter-symbol-interference (ISI) which degrades the SCM performance at high SE values. Equalizers with high computational complexity are required when the modulation bandwidth exceeds the 3-dB bandwidth of the VLC system.

Direct-current (DC)-biased optical OFDM (DCO-OFDM) is the most widely used multi-carrier modulation (MCM) technique for VLC. This section is focused on OFDM due to its potential in enabling high-speed VLC [12,20,56]. The achievable data rate of an OFDM-based VLC system can be maximized by adapting the system loading based on the frequency response of the considered channel. This can be achieved using the computationally efficient single-tap equalizers in addition to adaptive bit and energy loading algorithms. Adaptive bit loading is illustrated in Figure 31.6b. An improved throughput can be achieved by adaptively choosing the modulation order of M-QAM based on the achieved SNR at each subcarrier frequency. MCM modulation techniques have a relatively high peak to average power ratio (PAPR) compared with SCM modulation techniques due to the coincidence of multiple symbols with similar phase in the same frame. This can cause degradation in a VLC system performance when the dynamic range of the light source is limited. The peak and average values of an OFDM waveform are shown in Figure 31.6a. Colour modulation techniques allow for a new degree of freedom that



(a)

(b)

FIGURE 31.6 (a) The peak and average values of an OFDM waveform shown on the time-domain signal and the corresponding PDF. (b) The SNR of an OFDM-based VLC system with a violet m-LED [12]. Adaptively loaded bits are shown as a function of the achieved SNR at the respective subcarrier frequency.

is unique for VLC systems. In this section, several modulation techniques for VLC are reviewed.

31.4.1 Single-Carrier Modulation Techniques

Binary bits are encoded in the presence and absence of light intensity in OOK. Similarly multi-level intensity levels can encode the incoming bits with M-PAM [57]. Illumination control can be enabled by adjusting the light intensity of the two binary states in OOK without degrading the system performance. Compensation symbols are proposed in the VLC standard, IEEE 802.15.7 [58], to enable illumination control at the expense of reducing the SE. The position of the optical pulse is modulated into shorter duration pulses in PPM with a position index that varies depending on the incoming bits. PPM is more power efficient than OOK, but it requires more bandwidth to support equivalent data rates.

Performance comparisons between SCM and MCM techniques were considered in [59–61]. The performance of SCM techniques deteriorates as the data rates increase due to the increased ISI. SCM modulation techniques have less clipping distortion due to the lower PAPR compared with MCM modulation techniques. Carrier-less amplitude and phase (CAP) modulation was proposed as a solution to the high PAPR of MCM. The spectrum in CAP can be divided into m sub-carriers by the aid of multiple finite impulse response filters to enable adaptive loading of the VLC channel. High spectral efficiencies up to 10 bits (s Hz)$^{-1}$ were reported to be theoretically feasible at the expense of high computational complexity [62].

31.4.2 OFDM-Based MCM

SCM techniques require computationally complicated equalization processes when employed at high data rates due to the limited 3-dB bandwidth of the light source. In addition, effects such as the flickering interference of florescent lights may influence the system performance at the low frequency regions of the used bandwidth. On the other hand, MCM techniques such as OFDM can equalize the VLC channel by using the computationally efficient single-tap equalizers. Supporting multiple access is straightforward in MCM as it can be achieved by allocating distinctive subcarriers to each individual user terminal. In addition, OFDM supports adaptive bit and energy loading which allocates different modulation orders based on the estimated SNR at each individual subcarrier. This allows sub-carriers with high SNR values to be loaded with a higher modulation order. Therefore, OFDM can maximize the VLC system throughput by adapting the channel utilization to the low-pass frequency response of the VLC channel.

31.4.2.1 DCO-OFDM

Incoming bits are modulated into modulation formats such as M-PAM or M-QAM, where M is the modulation order given as $M = 2^b$ and b is the number of encoded bits in each M-PAM or M-QAM symbol. The modulated symbols are loaded afterwards into orthogonal subcarriers with a sub-carrier spacing equal to multiples of the symbol duration. The symbols can then be transformed into a time domain waveform using the inverse fast Fourier transformation (IFFT). The light intensity is real valued and strictly positive. Therefore, OFDM waveforms are required to be both unipolar and real valued. Hermitian symmetry is generally imposed on the frames of modulated symbols before an IFFT is performed to enforce the OFDM time domain output into the real domain. Different variants of optical OFDM were proposed to achieve a unipolar OFDM output.

DC bias is used in DCO-OFDM to shift the OFDM signal to the dynamic range of the light source [63]. However, OFDM signals have a high PAPR, which makes it practically impossible to convert all of the signal samples into unipolar ones. The OFDM time-domain waveform can be approximated with a Gaussian distribution when the length of the input frame is greater than 64 [64]. The DC bias point would be ideally within the centre of the linear region of the light source dynamic range. The dynamic range of a light source is the range in between the turn-on voltage and the voltage at the saturation power of the light source. OFDM values that fall outside the dynamic range incur clipping distortion. The optimization of the DC biasing point was studied in [65]. The DC bias can be defined as a k multiple of the standard deviation of the time-domain OFDM signal σ. The electrical power penalty due to the addition of DC bias in DCO-OFDM compared with bipolar OFDM can be written as:

$$B_{\mathrm{DC}}^{\mathrm{dB}} = 10\log_{10}\left(k^2+1\right). \tag{31.7}$$

The additional dissipation of electrical power caused by the DC bias in DCO-OFDM increases as the modulation order increases, which leads to electrical and optical power inefficiency. The electrical power penalty caused by the addition of DC bias is estimated in Figure 31.7 using Monte Carlo simulation based on (31.7) at a BER of 1×10^{-4}. The DCO-OFDM optical power inefficiency can be justified for some VLC applications where a high luminance is desired. However, an alternative modulation approach is required to achieve less energy consumption.

31.4.2.2 Unipolar OFDM Techniques

Unipolar OFDM modulation techniques were introduced to provide energy efficient optical OFDM alternatives to DCO-OFDM. These schemes include ACO-OFDM [66], pulse-amplitude-modulated discrete multi-tone modulation (PAM-DMT) [67], flipped OFDM (Flip-OFDM) [68] and unipolar orthogonal frequency division multiplexing (U-OFDM) [69]. The clipping distortion of an OFDM waveform at zero can be given as [11]:

$$x^{\mathrm{c}}[n] = f_{\mathrm{clip}}\left\{x[n]\right\} = \frac{1}{2}\left(x[n]+\left|x[n]\right|\right). \tag{31.8}$$

Unipolar OFDM modulation techniques exploit the OFDM input/output frame structure to produce a unipolar time domain waveform output. The modulated M-QAM symbols are loaded on the odd-indexed subcarriers in ACO-OFDM while the even-indexed sub-carriers are not utilized. This results in symmetry in the time domain ACO-OFDM waveform as shown in Figure 31.8a. In addition, it allows for the clipping distortion to fall on the even-indexed sub-carriers without affecting the information on the odd-indexed subcarriers as shown in Figure 31.8b [66]. The clipping operation given in (31.8) results in attenuation at the odd-indexed sub-carriers. However, skipping half of the sub-carriers reduces the SE of ACO-OFDM to half of that in DCO-OFDM. The imaginary component of the sub-carriers is loaded with M-PAM symbols in PAM-DMT while the real component is not utilized, which allows for the clipping distortion to affect the real component of the sub-carriers only as shown in Figure 31.9. PAM-DMT is more attractive than ACO-OFDM when bit-loading techniques are considered, as the PAM-DMT performance can be optimally adapted to the frequency response of the channel since all of the sub-carriers are used [67]. The SE of PAM-DMT is similar to that of DCO-OFDM. However, the bit error rate (BER) performance of M-PAM is similar to the performance of M^2-QAM. Therefore, PAM-DMT becomes energy inefficient as

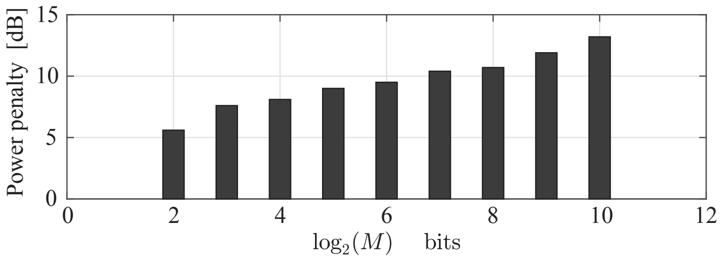


FIGURE 31.7 Power penalty caused by the addition of DC bias in DCO-OFDM compared with bipolar OFDM as a function of the modulation order based on (31.7).

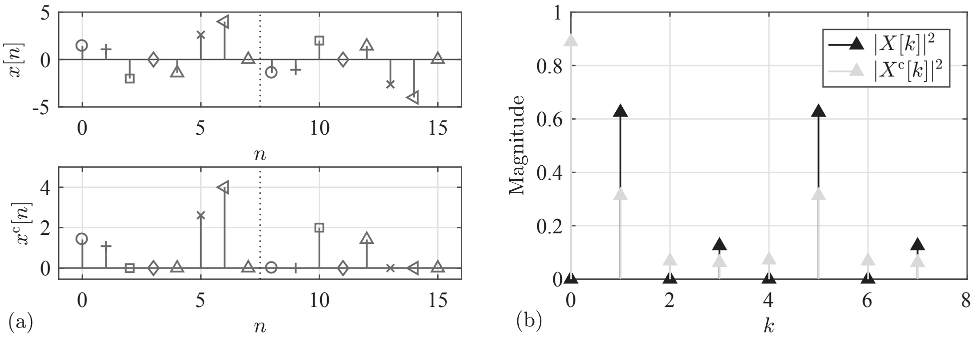


FIGURE 31.8 (a) Time domain ACO-OFDM waveform before $(x[n])$ and after clipping $(x^{c}[n])$. (b) Frequency domain representation of ACO-OFDM waveform before $(X[k])$ and after clipping $(X^{c}[k])$.

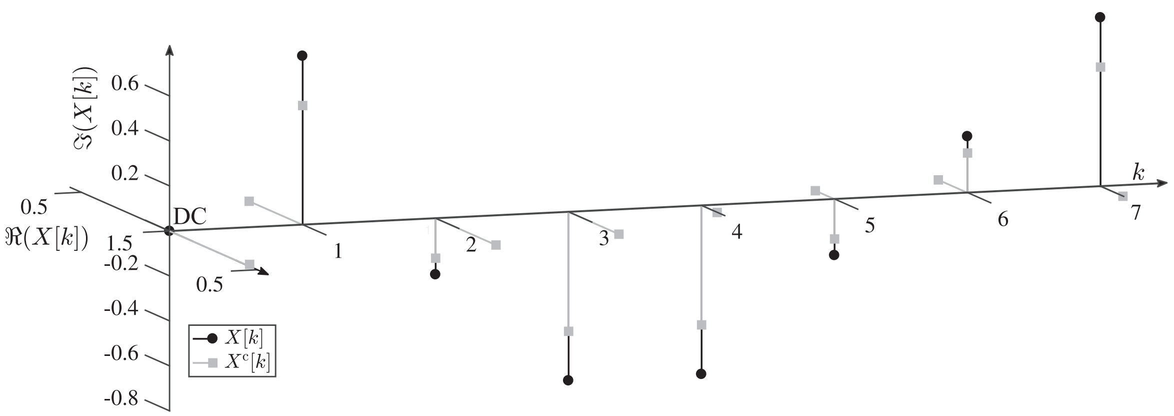


FIGURE 31.9 The frequency representation of the transmitted PAM-DMT waveform before $(X[k])$ and after clipping $(X^c[k])$.

the modulation order of M-PAM increases. Similar to ACO-OFDM, PAM-DMT has a 3 dB fixed penalty when compared with bipolar OFDM as half of the power is lost due to clipping. The bipolar time-domain OFDM frames are expanded into two unipolar frames in U-OFDM and Flip-OFDM. The first frame has the positive samples of the first original frame, while the second frame has a flipped version of the negative samples as illustrated in Figure 31.10. The concept and performance of U-OFDM and Flip-OFDM is identical [68,69]. The original frames can be reconstructed at the Rx by subtracting each second frame from the first frame in the same pair. This would double the noise at the Rx, which leads to a 3 dB penalty when compared with bipolar OFDM. The SE of U-OFDM is half of the SE of DCO-OFDM since two U-OFDM frames are required to convey the same information of a single DCO-OFDM frame. The reduced SE of ACO-OFDM, PAM-DMT and U-OFDM compared with DCO-OFDM requires these techniques to use a higher modulation order of M-PAM and M-QAM to match the SE of DCO-OFDM. Therefore, the BER performance of these unipolar techniques degrades at high modulation orders as they are more susceptible to noise.

31.4.2.3 Hybrid OFDM Techniques

Superposition-based enhanced U-OFDM (eU-OFDM) adds an additional degree of freedom to optical OFDM by allowing multiple U-OFDM streams to be transmitted simultaneously with the aid of successive interference cancellation (SIC) at the Rx [70]. This increases the SE of inherent unipolar techniques at the expense of increased computational and implementation complexities. Similar techniques with comparable performance were proposed based on ACO-OFDM [71,72] and PAM-DMT [73,74]. The OFDM sub-carriers are transmitted using different LEDs in the same array in spatial optical OFDM (SO-OFDM)[75]. SO-OFDM is reported to have BER performance gains over DCO-OFDM at high SNR due to the reduced PAPR and the robustness against LED non-linearities [75]. Reverse polarity optical OFDM (RPO-OFDM) was proposed to allow for a higher degree of illumination control by combining ACO-OFDM with a PWM [76].

31.4.3 Colour Modulation Techniques

Colour modulation techniques add a new degree of freedom that is unique to VLC. RGB-LEDs can produce different colours based on the intensity applied on each individual coloured-LED. The IEEE 802.15.7 standard proposes colour shift keying (CSK) for VLC [58]. The incoming bits are mapped into a constellation of colours from the chromatic CIE 1931 colour space [77]. The CIE 1931 is the illumination model for human eye colour perception. The incoming bits are modulated using the instantaneous colour of multi-colour LEDs. The overall intensity of the LED is constant which allows for constant optical power and colour as well as mitigating any flicker. Dimming can be achieved using a brightness control in CSK. However, colour shifts are possible due to the possible variations in the driving current. The Rx architecture has not been fully addressed in most of the published research on colour domain modulation. CSK is considered to be an expensive and complex modulation technique when compared with OFDM.

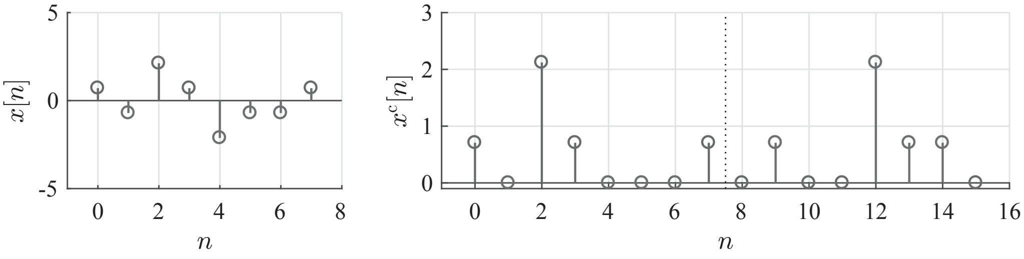


FIGURE 31.10 The generation of two U-OFDM frames from one bipolar OFDM frame.

31.5 Multiple-Input Multiple-Output Techniques in VLC and LiFi

In most indoor environments, multiple light sources are available to provide sufficient lighting functionality. A single LED chip is insufficient to provide enough power for a standard lighting device in general. The use of multiple lighting sources in an indoor environment provides opportunities to apply MIMO techniques to improve the performance of VLC or LiFi systems. In this section, a number of basic but powerful MIMO techniques are introduced.

31.5.1 Spatial Multiplexing

One of the most well-known MIMO techniques is recognized as spatial multiplexing (SMP) [78]. It can create multiple parallel channels, as shown in Figure 31.11a. By transmitting multiple data streams via different channels, the total transmission data rate can be significantly increased.

In SMP, inter-channel interference significantly degrades the quality of each parallel channel. Therefore, appropriate signal processing at both the Tx and Rx sides is required to mitigate the effect of inter-channel interference. This is known as pre-coding and post-processing. By calculating the singular value decomposition of the channel matrix, the optimal matrices for pre-coding and post-processing can be obtained. However, this method requires accurate channel information at both Tx and Rx sides. Furthermore, in an IM/DD-based system, additional constraints have to be fulfilled, such as limited dynamic range and non-negative signals. These conditions have led to extensive studies on the optimal pre-coding and post-processing designs under various practical constraints for SMP in VLC systems [79].

31.5.2 Repetition Coding

In practice, the channels between different pairs of optical sources and detectors are correlated, which is known as spatial correlation. The channel matrix condition number is frequently used to determine the 'quality' of the MIMO channel. If the ith singular value of the channel matrix is represented by σ_i, the condition number can be defined as $\max(\sigma_i)/\min(\sigma_i)$. In the case that the spatial correlation is weak, the channel matrix has a small condition number and independent robust parallel channels can be established by precoding. In the case that spatial correlation is strong, the MIMO channel matrix has a large condition number and it is not possible to form multiple robust parallel channels [80]. In this condition, it is more reasonable to let the multiple Txs to send the same data stream to increase the signal power that can be collected by the Rx and thereby increasing the probability that the transmitted signal is successfully decoded. This is known as repetition coding, as shown in Figure 31.11b. Studies have shown that the repetition coding is superior to some receive diversity scenarios as the intensity-modulated signal amplitude can be always superposed constructively [81].

31.5.3 Spatial Modulation

In some cases, the channel matrix condition number is small but not sufficient for SMP. In such a condition, spatial modulation (SM) is a suitable choice. In SM, each optical source is considered as a modulation constellation point [82,83]. Extra information bits are mapped to these spatial constellation points in parallel with the conventional amplitude/phase modulation (AMP) such as PAM, as shown in Figure 31.11c. Thus, transmission SE can be improved. Another benefit of SM is that only one optical source is activated at each time instant. Consequently, there is no need to implement an individual driving circuit for each optical source. Only one driving circuit in conjunction with a switch between the driving circuit and optical sources are sufficient to achieve the SM functionality. Therefore, the use of less hardware makes SM an energy efficient technique.

In the case that low system complexity is required, the AMP part can be removed and only the spatial signature is used to convey information, which is known as space shift keying [84]. If system complexity and power budget permit, SM can also be further extended to a generalized case known as generalized

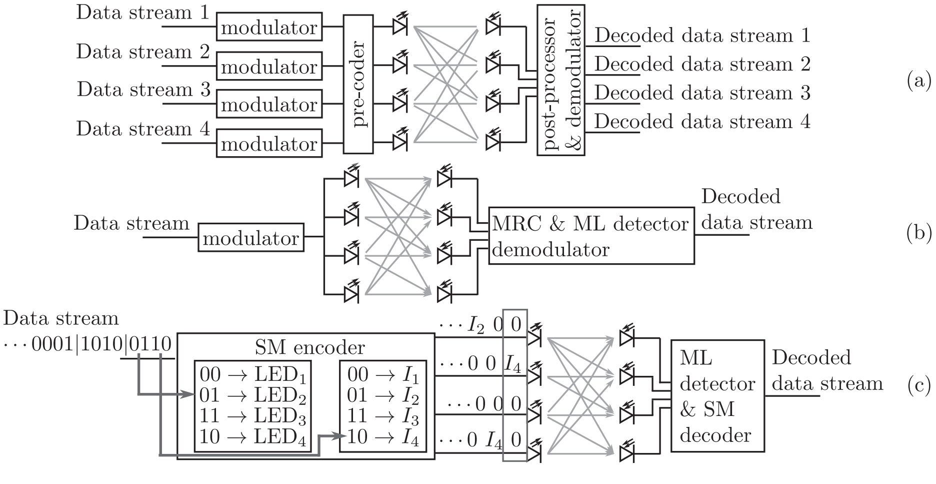

Data stream 1
Data stream 2
Data stream 3
Data stream 4
modulator
modulator
modulator
modulator
pre-coder
post-processor & demodulator
Decoded data stream 1
Decoded data stream 2
Decoded data stream 3
Decoded data stream 4
(a)
Data stream
modulator
MRC & ML detector demodulator
Decoded data stream
(b)
Data stream
$\cdots 0001|1010|0110$
SM encoder
$00 \rightarrow LED_1$
$01 \rightarrow LED_2$
$11 \rightarrow LED_3$
$10 \rightarrow LED_4$
$00 \rightarrow I_1$
$01 \rightarrow I_2$
$11 \rightarrow I_3$
$10 \rightarrow I_4$
$\cdots I_2\ 0\ 0$
$\cdots 0\ 0\ I_4$
$\cdots 0\ 0\ 0$
$\cdots 0\ I_4\ 0$
ML detector & SM decoder
Decoded data stream
(c)

FIGURE 31.11 Block diagram of MIMO techniques. (a) SMP; (b) repetition coding; (c) SM.

spatial modulation (GSM), which conveys information in various combinations of different active optical sources [85]. By using GSM, the same SE can be achieved with a fewer number of transmitting optical sources. Furthermore, additional spatial diversity can be explored.

31.5.4 Electrical-to-Optical Conversion Based on Multiple Light Sources

A typical optical source such as a LED has a non-linear forward current-to-optical power conversion characteristic. This issue limits the dynamic range of the optical source and causes distortion to the transmitted signal [10]. The use of multiple optical sources provides an opportunity to avoid the non-linearity issue [86]. First, the analog signal waveform is quantized and represented by a digital binary vector. The optical sources are divided into multiple groups and each optical source group can be switched on and off independently. Then, each transmitted signal sample can be represented by a certain light power level that is formed by switching on a particular combination of optical source groups which is controlled by the digital binary vectors after quantization. This approach is called discrete power level stepping. Since the optical power scales with the number of activated optical source groups, the non-linearity issue can be removed. In a following experimental study [87], discrete power level stepping is implemented using a CMOS-integrated GaN LED array. A data rate of 200 Mbps has been demonstrated using PAM. Furthermore, OFDM transmission has also been attempted. In addition to non-linearity, high PAPR is another major limiting factor to the performance of optical OFDM, as introduced in Section 31.4. It has been shown that by using multiple optical source groups, the OFDM high PAPR issue can be mitigated [75]. The time domain OFDM signal is the superposition of the low PAPR time domain signals on each OFDM sub-carrier. Therefore, it is possible to let each optical source group transmit the time domain signal on one of the OFDM subcarriers and the signals from each source groups are combined in free space. This method is also named as spatial summing [75].

31.5.5 MIMO Optical Source and Detector Deployment in VLC and LiFi

As introduced in Section 31.5.2, spatial correlation or channel matrix condition number is crucial to the performance of MIMO techniques. Therefore, it is important to deploy the transmitting optical sources and receiving detectors in a way that the spatial correlation is minimized. Note that the concepts of spatial correlation and channel matrix condition number have been introduced in Section 31.5.2. On the Tx side, multiple illumination devices or elements can be controlled and modulated individually, which form the multiple transmitting optical sources. These optical sources can be spatially separated. For example, using multiple separated lamps as MIMO transmitting elements with each lamp being treated as a single optical source [88], as shown in Figure 31.12a. In this case, the transmitting optical sources are spatially distributed. Considering a mobile user that is randomly located in a room, the light radiant angles and distances from every optical source to the Rx are likely to have different values. Consequently, the spatial correlation is weak and a channel matrix with a small condition number can be obtained. Alternatively, the transmitting optical sources can be integrated and close to each other in space. For example, the multiple LED modules on a single lamp can be treated as different MIMO transmitting elements [89], as shown in Figure 31.12b. In this case, the optical sources are close to each other, which means that the distances between the Rx and the optical sources are equal. Therefore, it is only possible to create difference in light radiant angle respect to each optical source by deploying optical sources with different orientations.

On the Rx side, a typical user equipment (UE) has a limited size. Consequently, it is not practical to increase the spatial separation between optical detectors for lower spatial correlation. Creating differences in incident angle respect to each optical detector is the only method to decrease the spatial correlation. A popular method is to deploy a single imaging lens above a detector array, as shown in Figure 31.12c. Each detector corresponds to a specific range of incident angle. In general, an imaging Rx is very compact in size. It offers a channel matrix with low spatial correlation and wide FoV [80]. However, in order to fit multiple detectors into a compact space, the physical area of each detector has to be small. In addition, the optical power from an optical source can only activate one or a few adjacent detectors in the array. These facts cause the low sensitivity of the imaging Rx. Another effective approach is to mount multiple detectors on the UE with different orientations. A number of designs following this method have been proposed [90,91], as shown in Figure 31.12d. Most of the mobile devices such as smart phone have a cubic shape, which makes the cubic Rx design (third design shown in Figure 31.12d) more convenient to be implemented in small mobile devices.

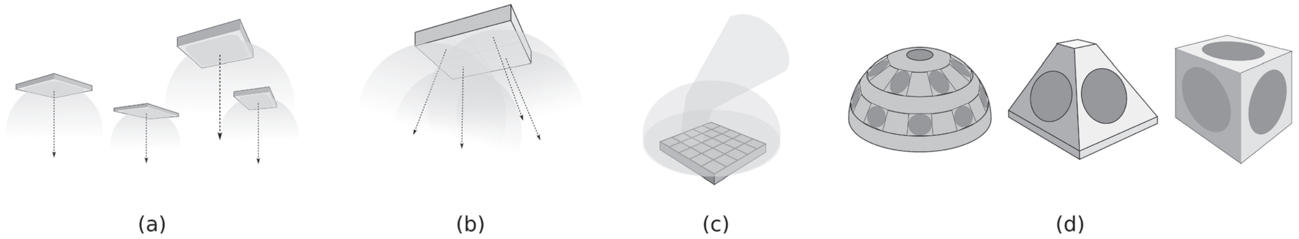


FIGURE 31.12 (a) Multiple transmitting optical sources formed by different lamps. (b) Multiple transmitting optical sources formed by different modules of a single lamp. (c) Optical detector array with imaging lens. (d) Optical detectors with various orientations.

31.5.6 Adaptation of MIMO Techniques

In practical VLC and LiFi systems, the VLC channel depends on the user location and the orientation of the mobile device. This channel uncertainty makes it inefficient to use a single MIMO technique in all conditions. When the channel matrix condition number is large, it is inefficient to use SMP to create parallel transmissions. Instead, RC is a more suitable candidate to enhance the SNR at the Rx side. When the channel matrix condition number is small, using RC may lead to insufficient transmission data rate or unnecessary power consumption for a target data rate. Therefore, SM/SMP can be selected for higher spectral or energy efficiency. This is the motivation of using adaptive MIMO algorithms in VLC and LiFi [92,93].

In an adaptive MIMO algorithm, first, the channel state is evaluated. Then, the performance metric is estimated with various MIMO techniques. The one that achieves the highest performance metric is selected. For users with movement, this adaptive algorithm has to be repeated periodically. The time between two adaptations has to be equal or smaller than the coherence time of the channel. Coherence time is defined as the time over which the change of the channel is negligible. In a high performance LiFi network, achievable data rate or SE is a suitable metric. The adaptive algorithm with maximization of achievable data rate has been proposed in [92]. Energy efficiency is another commonly used performance metric in applications such as the IoT. The adaptive algorithm based on this metric has been developed in [93].

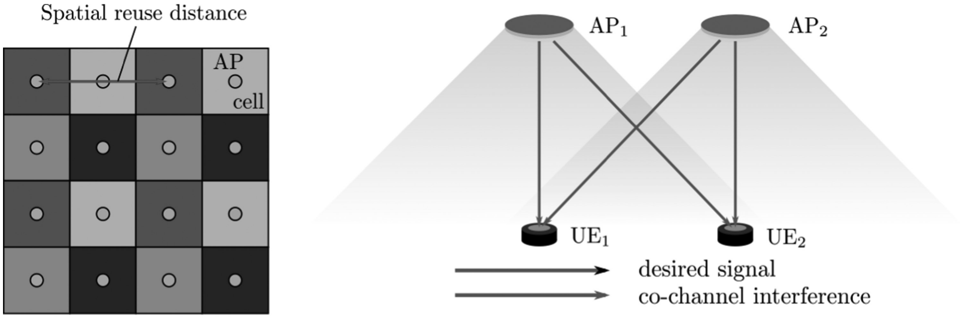
31.6 Co-channel Interference and Interference Mitigation Techniques in Multi-Cell LiFi Networks

Multiple illumination sources in a LiFi network can provide not only large coverage areas but also the opportunity to implement a network as a cellular system which is named as a LiFi attocell network. This is similar to the topology of a conventional cellular network in RF wireless systems, where multiple APs are deployed and each AP serves a number of nearby mobile users. The region that an AP covers is defined as a cell. In a LiFi attocell network, the limited transmission resources, such as time, frequency and code, can be reused by APs that are spatially separated. By doing so, the achievable bit rate per unit bandwidth, known as SE, of the entire LiFi attocell network can be significantly improved. Considering the attocell network shown in Figure 31.13a, the transmission resource has to be divided into 16 pieces for the signal transmission in 16 cells if there is no spatial reuse. In contrast, each cell is able to use one-fourth of the total transmission resource if the spatial reuse pattern shown in the figure is deployed, thereby improving the network SE by a factor of four. The shortest distance between cells reusing the same transmission resource is defined as the spatial reuse distance. On the other hand, different wireless transmissions using the same transmission resource cause interference to each other. This is called co-channel interference (CCI), as demonstrated in Figure 31.13b. CCI causes an extra distortion to the desired signal in addition to the effect of noise. With a decrease in reuse distance, the distortion caused by CCI increases.

31.6.1 Access Point Layout and SINR Characterizations

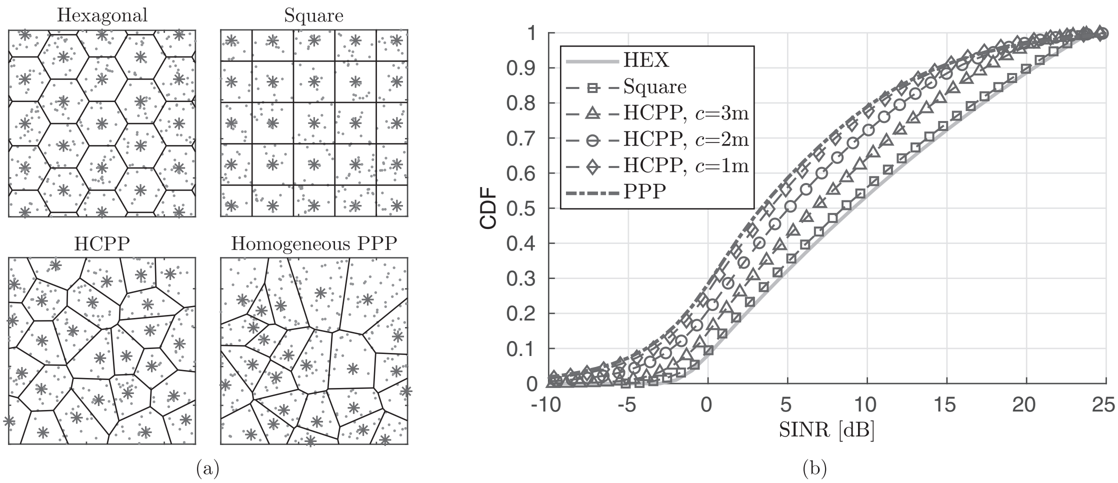
The aggregate interference is the accumulation of CCI from multiple nearby APs. One would intuitively induce that the aggregate interference statistics would converge to a Gaussian distribution with the increase in the number of interfering APs due to the law of large number. However, this is not the case as the aggregate interference is always dominated by several of the closest interfering APs due to the near-far field issue [94]. This makes the statistics of the aggregate interference significantly depend on the AP deployments of the LiFi network. It has been found in the RF cellular system studies that the performance of actual cellular systems can be upper bounded by that of a hexagonal grid-based base station (BS) deployment and lower bounded by that of a random point process-based BS deployment [95]. In order to characterize the CCI in a LiFi network, the same methodology has been adopted.

It is desired to make an AP have a circular coverage area. This is because of the central symmetric radiation pattern of the optical Tx. However, it is not possible to make such circular coverage with Voronoi tessellation to form the non-overlapping cell coverage by each AP. An optimum hexagonal AP layout is shown in Figure 31.14a. With this cell layout, the separation between the coverage area of an AP and the nearby APs are maximized, which leads to the minimum level of CCI. In the other extreme case, the AP layout follows a two-dimensional



(a) (b)

FIGURE 31.13 (a) Spatial reuse and reuse distance. (b) CCI in LiFi attocell network.



(a) (b)

FIGURE 31.14 (a) AP layout in LiFi networks. Top left: hexagonal grid layout; top right: square grid layout; bottom left: HCPP layout; bottom right: homogeneous PPP layout. (b) SINR statistics with various AP layouts. In the case of HCPP, c means the minimum repulsive distance between APs.

homogeneous Poisson point process (PPP). With this AP layout, the position of each AP is completely random. Therefore, the coverage area of each AP is an irregular polygon of various sizes and numbers of edges. Furthermore, there is no minimum distance between the coverage area of an AP and the nearby APs. Consequently, the level of CCI is expected to be very high in a network with this AP layout. These two AP layouts represent the best and worst case configurations. However, the hexagonal and PPP luminary layout is less common in an illumination network. Therefore, it is possible but not likely to configure a network following these AP layouts in practice. Figure 31.14a also demonstrates another two AP layouts that are more practical. The former places APs on a square lattice, and the latter distributes the APs with a repulsive point process (RPP). The layout with square lattice is the most common one in daily life. On the other hand, the RPP takes random AP placement into account, but a minimum separation between each pair of APs is required. In this section, we use an AP layout with hard-core point process (HCPP) as an example. The AP layout randomness is motivated by the following observations: in some application scenarios, lighting uniformity is not required, and the lighting effect is enhanced in certain specified locations; in a network with a regular AP layout, some of the APs may be in silent mode because there is no user to serve. It is shown in Figure 31.14 that the performance of networks in terms of signal-to-interference-plus-noise ratio (SINR) with a square grid and HCPP AP layout is bounded by the two extreme cases with hexagonal and PPP AP layouts. A number of studies on the characterization of CCI and SINR based hexagonal and PPP AP layout models have been conducted [5,96,97].

31.6.2 Interference Mitigation Techniques

By reducing the spatial reuse distance, the SE of a LiFi attocell network can be improved. However, the introduced CCI makes the mobile users have a probability of over 50% to receive a SINR lower than 10 dB, as shown in Figure 31.14b. In this case, various transmission resources can be intelligently allocated to different cells and users to avoid or mitigate the negative effects of CCI. These are known as interference mitigation techniques.

A simple and effective method to avoid CCI is to assign different transmission resources to users in adjacent cells, which is known as resource partitioning [98,99]. Significant reduction in SE is the main disadvantage of this approach. A variation of the resource partitioning method is known as fractional frequency reuse (FFR) [100]. It assigns transmission resources with short spatial reuse distance to users in cell centre regions, as these users are generally limited by noise. Transmission resources with longer spatial reuse distance are assigned to UEs in cell edge regions which are generally suffering from CCI with the power that is comparable to the desired signal power. FFR can achieve a good balance between CCI mitigation capability and SE with relatively low complexity. However, transmission resources may be wasted, especially in a lightly loaded network. Another unique way of removing the negative effects of CCI is by converting harmful interference into the useful designed signal. This idea can be realized by a technique known as coordinated multipoint joint transmission (CoMP-JT) used in RF cellular systems [101]. A very important advantage of applying CoMP-JT in a LiFi network is that the intensity-modulated signals can always be combined constructively without accurate alignment or synchronization [102]. This characteristic makes the implementation of CoMP-JT much simpler in a LiFi attocell network than in a RF cellular network. A more efficient approach with higher complexity to mitigate CCI has been proposed in [103]. A user-centric cluster is formed by multiple closely located users. The users in the cluster are jointly served by multiple nearby APs. With MIMO precoding techniques following the zero-forcing criterion, the data for each user within the cluster can be transmitted without interfering each other using vector transmission. Compared to CoMP-JT, the technique using user-centric cluster formation and vector transmission is able to adapt the network based on its load condition. However, with an increase of number of users, the required time and computational resources for channel estimation, cluster formation and pre-coder calculation

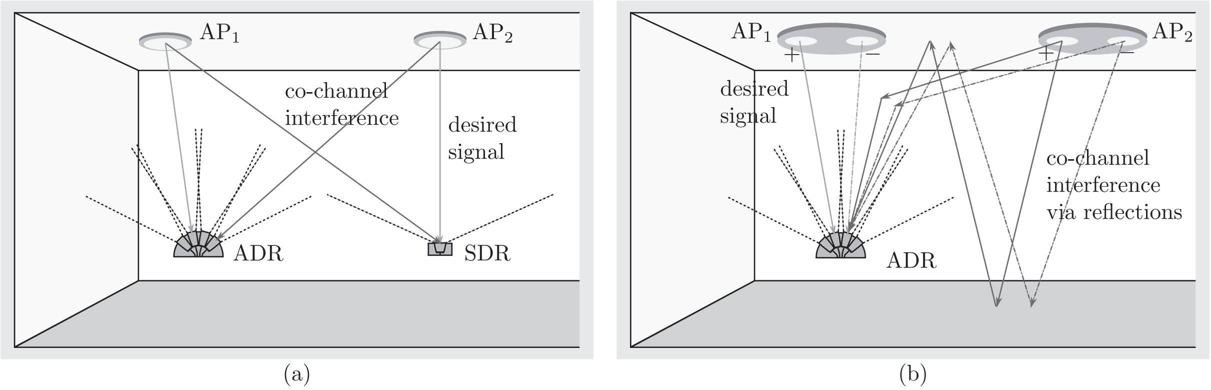


FIGURE 31.15 (a) CCI mitigation with ADR. (b) CCI mitigation with ADR and DST.

will increase significantly. Therefore, it is more suitable to be applied in a lightly loaded LiFi network.

The aforementioned interference mitigation techniques avoid interference by intelligently assigning transmission resources. In contrast, the following method removes the effects of CCI by exploring angular diversity at the Tx and the Rx. In a LiFi network, the multiple optical sources are likely to be separated in the room. This fact makes signals from multiple optical sources incident to the user device Rx from different spatial angles. Taking advantage of this characteristic, we can integrate multiple detecting elements to a single Rx with various orientations, as shown in Figure 31.15a. By using this angular diversity Rx (ADR), unwanted interfering signals can be rejected without sacrificing the desired signal quality [34]. This is similar to a simplified receive beam-forming system. However, there is no need to use complex signal processing to form the 'beam', as each single optical detector element is readily directional. On top of the usage of ADR, a double-source Tx (DST) can be introduced to further explore the angular diversity on the Tx side [104]. The design is illustrated in Figure 31.15b. Two visible light sources are deployed side-by-side. Both sources transmit the same signal with an opposite polarity. In the desired coverage area underneath the Tx, the two signal streams are likely to be collected by different detecting elements on the Rx. Then the two signal streams can be appropriately combined using advanced combining schemes. In the area further away from the Tx, the angle of incidence is similar, and the two signal streams are more likely to be captured by the same detecting elements. Consequently, the two signal streams are added destructively and the interference to the undesired users can be significantly mitigated. This method is found to be very effective in cancelling CCI via the NLoS channels.

31.7 Conclusion

In the future, ubiquitous connectivity will hold an even greater importance in people's daily lives. How to realize fast, secure wireless access in dense user scenarios is one of the major challenges for future wireless network design. High-performance, multi-functional LiFi networks are prominent candidates to fulfil this requirement. By exploring the spatial characteristics of optical signals, the LiFi network is able to deliver high-speed secure wireless internet access for a large number of mobile users. Furthermore, additional functions such as indoor positioning and connectivity for IoT are also potential applications for LiFi technology.

Extensive physical layer research in the past few years has made high speed VLC and LiFi networks possible. In particular, advances in the modulation bandwidths of LEDs and LDs have enabled high speed wireless transmission with a data rate of several giga-bits per second. Various novel modulation techniques that are suitable for VLC have been developed to achieve improved energy and SE with the illumination requirement fulfilled. It has been shown that MIMO techniques have the potential to enhance data rates, but they require increased hardware complexity at the Tx and the Rx as well as increased computational complexity. This chapter has also showcased new channel modelling approaches for mobile networks based on LiFi which enable seamless coverage to support user mobility. LiFi requires multiple optical APs causing interference. Therefore, interference mitigation techniques have been introduced to ensure high data densities for the entire coverage area.

REFERENCES

1. Cisco Visual Networking Index, Global Mobile Data Traffic Forecast Update, 2016-2021, White Paper, CISCO (Jun. 2017).
2. H. Haas, L. Yin, Y. Wang, C. Chen, What is LiFi? *J. Lightwave Technol*. 34 (6) (2016) 1533–1544. doi:10.1109/JLT.2015.2510021.
3. M. D. Soltani, X. Wu, M. Safari, H. Haas, Bidirectional user throughput maximization based on feedback reduction in LiFi networks, *IEEE Trans. Commun*. (2018) 1–1 doi:10.1109/TCOMM.2018.2809435.
4. L. Yin, H. Haas, Physical-layer security in multiuser visible light communication networks, *IEEE J. Sel. Areas Commun*. 36 (1) (2018) 162–174. doi:10.1109/JSAC.2017.2774429.
5. C. Chen, D. A. Basnayaka, H. Haas, Downlink performance of optical attocell networks, *J. Lightwave Technol*. 34 (1) (2016) 137–156. doi:10.1109/JLT.2015.2511015.

6. J. Brodrick, Suggested Research Topics Supplement: Technology and Market Context, Tech. Rep., DOE SSL Program (2017).
7. G. Stepniak, J. Siuzdak, P. Zwierko, Compensation of a VLC phosphorescent white LED nonlinearity by means of Volterra DFE, *IEEE Photon. Technol. Lett.* 25 (16) (2013) 1597–1600. doi:10.1109/LPT.2013.2272511.
8. W. Zhao, Q. Guo, J. Tong, J. Xi, Y. Yu, P. Niu, X. Sun, Orthogonal polynomial-based nonlinearity modeling and mitigation for LED communications, *IEEE Photon. J.* 8 (4) (2016) 1–12. doi:10.1109/JPHOT.2016.2581485.
9. B. Inan, S. C. J. Lee, S. Randel, I. Neokosmidis, A. M. J. Koonen, J. W. Walewski, Impact of LED nonlinearity on discrete multitone modulation, *IEEE/OSA J. Opt. Commun. Netw.* 1 (5) (2009) 439–451. doi:10.1364/JOCN.1.000439.
10. H. Elgala, R. Mesleh, H. Haas, Non-linearity effects and predistortion in optical OFDM wireless transmission using LEDs, *Int. J. Ultra Wideband Commun. Syst.* 1 (2) (2009) 143–150. doi:10.1504/IJUWBCS.2009.029003.
11. D. Tsonev, S. Sinanovic, H. Haas, Complete modeling of nonlinear distortion in OFDM-based optical wireless communication, *J. Lightwave Technol.* 31 (18) (2013) 3064–3076. doi:10.1109/JLT.2013.2278675.
12. M. S. Islim, R. X. Ferreira, X. He, E. Xie, S. Videv, S. Viola, S. Watson, N. Bamiedakis, R. V. Penty, I. H. White, A. E. Kelly, E. Gu, H. Haas, M. D. Dawson, Towards 10 Gb/s orthogonal frequency division multiplexing-based visible light communication using a GaN violet micro-LED, *Photon. Res.* 5 (2) (2017) A35–A43. doi:10.1364/PRJ.5.000A35.
13. S. Dimitrov, H. Haas, *Principles of LED Light Communications: Towards Networked Li-Fi*, Cambridge University Press, 2015.
14. A. M. Khalid, G. Cossu, R. Corsini, P. Choudhury, E. Ciaramella, 1-Gb/s transmission over a phosphorescent white LED by using rate-adaptive discrete multitone modulation, *IEEE Photon. J.* 4 (5) (2012) 1465–1473. doi:10.1109/JPHOT.2012.2210397.
15. Y. Zhou, J. Zhao, M. Zhang, J. Shi, N. Chi, 2.32 Gbit/s phosphorescent white LED visible light communication aided by two-staged linear software equalizer, in: *2016 10th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, 2016, pp. 1–4. doi:10.1109/CSNDSP.2016.7573913.
16. C. Foucher, M. I. Sufyan, B. J. E. Guilhabert, S. Videv, S. Rajbhandari, A. G. Diaz, H. Chun, D. A. Vithanage, G. A. Turnbull, I. D. W. Samuel, G. Faulkner, D. C. O'Brien, H. Haas, N. Laurand, M. D. Dawson, Flexible glass hybridized colloidal quantum dots for Gb/s visible light communications, *IEEE Photon. J.* 10 (1) (2018) 1–11. doi:10.1109/JPHOT.2018.2792700.
17. M. F. Leitão, J. M. M. Santos, B. Guilhabert, S. Watson, A. E. Kelly, M. S. Islim, H. Haas, M. D. Dawson, N. Laurand, Gb/s visible light communications with colloidal quantum dot color converters, *IEEE J. Sel. Topics. Quantum Electron.* 23 (5) (2017) 1–10. doi:10.1109/JSTQE.2017.2690833.
18. A. Neumann, J. J. Wierer, W. Davis, Y. Ohno, S. R. J. Brueck, J. Tsao, Four-color laser white illuminant demonstrating high color-rendering quality, *Opt. Express* 19 (S4) (2011) A982–A990. doi:10.1364/OE.19.00A982.
19. C. Lee, C. Shen, C. Cozzan, R. M. Farrell, S. Nakamura, A. Y. Alyamani, B. S. Ooi, J. E. Bowers, S. P. DenBaars, J. Speck, Semipolar GaN-based laser diodes for Gbit/s white lighting communication: devices to systems, in: *Proceedings of SPIE*, Vol. 10532, 2018, pp. 10532–10532–11. doi:10.1117/12.2315791.
20. S. Viola, M. S. Islim, S. Watson, S. Videv, H. Haas, A. E. Kelly, 15 Gb/s OFDM-based VLC using direct modulation of 450 GaN laser diode, in: *Proceedings of SPIE*, Vol. 10437, 2017, pp. 10437–10437–7. doi:10.1117/12.2292004.
21. J. M. Kahn, J. R. Barry, Wireless infrared communications, *Proc. IEEE* 85 (2) (1997) 265–298. doi:10.1109/5.554222.
22. D. Tsonev, S. Videv, H. Haas, Towards a 100 Gb/s visible light wireless access network, *Opt. Express* 23 (2) (2015) 1627–1637. doi:10.1364/OE.23.001627.
23. L.-Y. Wei, C.-W. Hsu, C.-W. Chow, C.-H. Yeh, 20.231 Gbit/s tricolor red/green/blue laser diode based bidirectional signal remodulation visible-light communication system, *Photon. Res.* 6 (5) (2018) 422–426. doi:10.1364/PRJ.6.000422.
24. Y.-C. Chi, D.-H. Hsieh, C.-Y. Lin, H.-Y. Chen, C.-Y. Huang, J.-H. He, B. Ooi, S. P. DenBaars, S. Nakamura, H.-C. Kuo, et al., Phosphorous diffuser diverged blue laser diode for indoor lighting and communication, *Scientific Reports* 5 (2015) 18690. doi:10.1038/srep18690.
25. I. Dursun, C. Shen, M. R. Parida, J. Pan, S. P. Sarmah, D. Priante, N. Alyami, J. Liu, M. I. Saidaminov, M. S. Alias, A. L. Abdelhady, T. K. Ng, O. F. Mohammed, B. S. Ooi, O. M. Bakr, Perovskite Nanocrystals as a Color Converter for Visible Light Communication, *ACS Photon.* 3 (7) (2016) 1150–1156. arXiv: https://doi.org/10.1021/acsphotonics.6b00187, doi:10.1021/acsphotonics.6b00187.
26. C. Lee, C. Shen, C. Cozzan, R. M. Farrell, J. S. Speck, S. Nakamura, B. S. Ooi, S. P. DenBaars, Gigabit-per-second white light-based visible light communication using near-ultraviolet laser diode and red-, green-, and blue-emitting phosphors, *Opt. Express* 25 (15) (2017) 17480–17487. doi:10.1364/OE.25.017480.
27. R. X. G. Ferreira, E. Xie, J. J. D. McKendry, S. Rajbhandari, H. Chun, G. Faulkner, S. Watson, A. E. Kelly, E. Gu, R. V. Penty, I. H. White, D. C. O'Brien, M. D. Dawson, High bandwidth GaN-based micro-LEDs for Multi-Gb/s visible light communications, *IEEE Photon. Technol. Lett.* 28 (19) (2016) 2023–2026. doi:10.1109/LPT.2016.2581318.
28. H. Chun, A. Gomez, G. Faulkner, D. O'Brien, A spectrally efficient equalization technique for optical sources with direct modulation, *Opt. Lett.* 43 (11) (2018) 2708–2711. doi:10.1364/OL.43.002708.
29. P. A. Haigh, Z. Ghassemlooy, S. Rajbhandari, I. Papakonstantinou, Visible light communications using organic light emitting diodes, *IEEE Commun. Mag.* 51 (8) (2013) 148–154. doi:10.1109/MCOM.2013.6576353.
30. T. Fukuda, T. Okada, B. Wei, M. Ichikawa, Y. Taniguchi, Influence of carrier-injection efficiency on modulation rate of organic light source, *Opt. Lett.* 32 (13) (2007) 1905–1907. doi:10.1364/OL.32.001905.
31. E. Sarbazi, M. Safari, H. Haas, Statistical modeling of single-photon avalanche diode receivers for optical wireless communications, *IEEE Transactions on Communications* (2018) 1–1 doi:10.1109/TCOMM.2018.2822815.

32. Y. Li, M. Safari, R. Henderson, H. Haas, Optical OFDM with single-photon avalanche diode, *IEEE Photonics Technology Letters* 27 (9) (2015) 943–946. doi:10.1109/LPT.2015.2402151.
33. Z. Wang, D. Tsonev, S. Videv, H. Haas, On the design of a solar-panel receiver for optical wireless communications with simultaneous energy harvesting, *IEEE J. Sel. Areas Commun.* 33 (8) (2015) 1612–1623. doi:10.1109/JSAC.2015.2391811.
34. Z. Chen, N. Serafimovski, H. Haas, Angle diversity for an indoor cellular visible light communication system, in: *Proceedings of IEEE Vehicular Technology Conference*, Seoul, South Korea, 2014, pp. 1–5. doi:10.1109/VTCSpring.2014.7022832.
35. J. B. Carruthers, J. M. Kahn, Angle diversity for nondirected wireless infrared communication, *IEEE Trans. Commun.* 48 (6) (2000) 960–969. doi:10.1109/26.848557.
36. M. S. Islim, S. Videv, M. Safari, E. Xie, J. J. D. McKendry, E. Gu, M. D. Dawson, H. Haas, The impact of solar irradiance on visible light communications, *J. Lightwave Technol.* 36 (12) (2018) 2376–2386.
37. P. P. Manousiadis, S. Rajbhandari, R. Mulyawan, D. A. Vithanage, H. Chun, G. Faulkner, D. C. O'Brien, G. A. Turnbull, S. Collins, I. D. Samuel, Wide field-of-view fluorescent antenna for visible light communications beyond the étendue limit, *Optica* 3 (7) (2016) 702–706. doi:10.1364/OPTICA.3.000702.
38. T. Peyronel, K. J. Quirk, S. C. Wang, T. G. Tiecke, Luminescent detector for free-space optical communication, *Optica* 3 (7) (2016) 787–792. doi:10.1364/OPTICA.3.000787.
39. S. Cova, M. Ghioni, A. Lacaita, C. Samori, F. Zappa, Avalanche photodiodes and quenching circuits for single-photon detection, *Appl. Opt.* 35 (12) (1996) 1956–1976. doi:10.1364/AO.35.001956.
40. Y. Li, S. Videv, M. Abdallah, K. Qaraqe, M. Uysal, H. Haas, Single photon avalanche diode (SPAD) VLC system and application to downhole monitoring, in: *2014 IEEE Global Communications Conference*, 2014, pp. 2108–2113. doi:10.1109/GLOCOM.2014.7037119.
41. D. Chitnis, L. Zhang, H. Chun, S. Rajbhandari, G. Faulkner, D. O'Brien, S. Collins, A 200 Mb/s VLC demonstration with a SPAD based receiver, in: *2015 IEEE Summer Topicals Meeting Series (SUM)*, 2015, pp. 226–227. doi:10.1109/PHOSST.2015.7248280.
42. S.-M. Kim, J.-S. Won, Simultaneous reception of visible light communication and optical energy using a solar cell receiver, in: *2013 International Conference on ICT Convergence (ICTC)*, 2013, pp. 896–897. doi:10.1109/ICTC.2013.6675511.
43. J. Fakidis, S. Videv, H. Helmers, H. Haas, 0.5-Gb/s OFDM-based laser data and power transfer using a GaAs photovoltaic cell, *IEEE Photon. Technol. Lett.* 30 (9) (2018) 841–844. doi:10.1109/LPT.2018.2815273.
44. S. Zhang, D. Tsonev, S. Videv, S. Ghosh, G. A. Turnbull, I. D. W. Samuel, H. Haas, Organic solar cells as high-speed data detectors for visible light communication, *Optica* 2 (7) (2015) 607–610. doi:10.1364/OPTICA.2.000607.
45. R. Zhang, C. K. Ho, MIMO broadcasting for simultaneous wireless information and power transfer, *IEEE Transactions on Wireless Communications* 12 (5) (2013) 1989–2001. doi:10.1109/TWC.2013.031813.120224.
46. J. Barry, J. Kahn, W. Krause, E. Lee, D. Messerschmitt, simulation of multipath impulse response for indoor wireless optical channels, *IEEE J. Sel. Areas Commun.* 11 (3) (1993) 367–379. doi:10.1109/49.219552.
47. T. Komine, M. Nakagawa, Fundamental analysis for visible-light communication system using LED lights, *IEEE Trans. Consum. Electron.* 50 (1) (2004) 100–107. doi:10.1109/TCE.2004.1277847.
48. J. B. Carruthers, P. Kannan, Iterative site-based modeling for wireless infrared channels, *IEEE Transactions on Antennas and Propagation* 50 (5) (2002) 759–765. doi:10.1109/TAP.2002.1011244.
49. K. Lee, H. Park, J. Barry, Indoor channel characteristics for visible light communications, *IEEE Commun. Lett.* 15 (2) (2011) 217–219.
50. F. J. Lopez-Hernandez, R. Perez-Jimenez, A. Santamaria, Ray-tracing algorithms for fast calculation of the channel impulse response on diffuse IR wireless indoor channels, *Optical Eng.* 39 (10) (2000) 2775–2780. doi:10.1117/1.1287397.
51. O. Gonzalez, C. Militello, S. Rodriguez, R. Prez-Jimenez, A. Ayala, Error estimation of the impulse response on diffuse wireless infrared indoor channels using a Monte Carlo ray-tracing algorithm, *IEE Proceedings - Optoelectronics* 149 (56) (2002) 222–227. doi:10.1049/ip-opt:20020545.
52. H. Schulze, Frequency-domain simulation of the indoor wireless optical communication channel, *IEEE Trans. Commun.* 64 (6) (2016) 2551–2562. doi:10.1109/TCOMM.2016.2556684.
53. V. Jungnickel, V. Pohl, S. Nonnig, C. von Helmolt, A physical model of the wireless infrared communication channel, *IEEE J. Sel. Areas Commun.* 20 (3) (2002) 631–640. doi:10.1109/49.995522.
54. C. Chen, D. A. Basnayaka, X. Wu, H. Haas, Efficient analytical calculation of non-line-of-sight channel impulse response in visible light communications, *J. Lightw. Technol.* 36 (9) (2018) 1666–1682. doi:10.1109/JLT.2017.2786459.
55. J. B. Carruthers, J. M. Kahn, Modeling of nondirected wireless infrared channels, *IEEE Trans. Commun.* 45 (10) (1997) 1260–1268. doi:10.1109/26.634690.
56. D. Tsonev, H. Chun, S. Rajbhandari, J. J. D. McKendry, S. Videv, E. Gu, M. Haji, S. Watson, A. E. Kelly, G. Faulkner, M. D. Dawson, H. Haas and D. O'Brien, A 3-Gb/s single-LED OFDM-based wireless VLC link using a gallium nitride μLED, *IEEE Photon. Technol. Lett.* 26 (7) (2014) 637–640. doi:10.1109/LPT.2013.2297621.
57. S. Randel, F. Breyer, S. C. J. Lee, J. W. Walewski, Advanced modulation schemes for short-range optical communications, *IEEE J. Sel. Topics. Quantum Electron.* PP (99) (2010) 1–10. doi:10.1109/JSTQE.2010.2040808.
58. IEEE Std. 802.15.7-2011, IEEE Standard for Local and Metropolitan Area Networks, Part 15.7: Short-Range Wireless Optical Communication Using Visible Light (2011).
59. J. Armstrong, B. J. C. Schmidt, Comparison of asymmetrically clipped optical OFDM and DC-biased optical OFDM in AWGN, *IEEE Commun. Lett.* 12 (5) (2008) 343–345. doi:10.1109/LCOMM.2008.080193.
60. M. Kashani, M. Kavehrad, On the performance of single- and multi-carrie modulation schemes for indoor visible light communication systems, in: *Global Communications Conference (GLOBECOM)*, 2014 IEEE, 2014, pp. 2084–2089. doi:10.1109/GLOCOM.2014.7037115.

61. S. H. Lee, S.-Y. Jung, J. K. Kwon, Modulation and coding for dimmable visible light communication, *IEEE Commun. Mag.* 53 (2) (2015) 136–143. doi:10.1109/MCOM.2015.7045402.
62. P. Haigh, S. T. Le, S. Zvanovec, Z. Ghassemlooy, P. Luo, T. Xu, P. Chvojka, T. Kanesan, E. Giacoumidis, P. Canyelles-Pericas, H. L. Minh, W. Popoola, S. Rajbhandari, I. Papakonstantinou, I. Darwazeh, Multi-band carrier-less amplitude and phase modulation for bandlimited visible light communications systems, *IEEE Wireless Commun* 22 (2) (2015) 46–53. doi:10.1109/MWC.2015.7096284.
63. J. B. Carruthers, J. M. Kahn, Multiple-subcarrier modulation for nondirected wireless infrared communication, *IEEE J. Sel. Areas Commun.* 14 (3) (1996) 538–546. doi:10.1109/49.490239.
64. M. S. Islim, H. Haas, Modulation techniques for Li-Fi, *ZTE Commun.* 14 (2) (2016) 29–40.
65. S. Dimitrov, H. Haas, Information rate of OFDM-based optical wireless communication systems with nonlinear distortion, *J. Lightwave Technol.* 31 (6) (2013) 918–929. doi:10.1109/JLT.2012.2236642.
66. J. Armstrong, A. Lowery, Power efficient optical OFDM, *Electronics Letters* 42 (6) (2006) 370–372. doi:10.1049/el:20063636.
67. S. C. J. Lee, S. Randel, F. Breyer, A. M. J. Koonen, PAM-DMT for intensity-modulated and direct-detection optical communication systems, *IEEE Photonics Technology Letters* 21 (23) (2009) 1749–1751. doi:10.1109/LPT.2009.2032663.
68. N. Fernando, Y. Hong, E. Viterbo, Flip-OFDM for unipolar communication systems, *IEEE Transactions on Communications* 60 (12) (2012) 3726–3733. doi:10.1109/TCOMM.2012.082712.110812.
69. D. Tsonev, S. Sinanović, H. Haas, Novel unipolar orthogonal frequency division multiplexing (U-OFDM) for optical wireless, in: *Proceedings of the Vehicular Technology Conference (VTC Spring), IEEE*, IEEE, Yokohama, Japan, 2012.
70. D. Tsonev, S. Videv, H. Haas, Unlocking spectral efficiency in intensity modulation and direct detection systems, *IEEE J. Sel. Areas Commun.* 33 (99) (2015) 1758–1770. doi:10.1109/JSAC.2015.2432530.
71. M. Islim, D. Tsonev, H. Haas, On the superposition modulation for OFDM-based optical wireless communication, in: *IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE*, Orlando, USA, 2015.
72. Q. Wang, C. Qian, X. Guo, Z. Wang, D. G. Cunningham, I. H. White, Layered ACO-OFDM for intensity-modulated direct-detection optical wireless transmission, *Opt. Express* 23 (9) (2015) 12382–12393. doi:10.1364/OE.23.012382.
73. M. Islim, D. Tsonev, H. Haas, Spectrally enhanced PAM-DMT for IM/DD optical wireless communications, in: *Proceedings of the 25th International Symposium. Personal, Indoor and Mobile Radio Communication (PIMRC)*, IEEE, Hong Kong, China, 2015, pp. 927–932.
74. M. S. Islim, H. Haas, Augmenting the spectral efficiency of enhanced PAM-DMT-based optical wireless communications, *Opt. Express* 24 (11) (2016) 11932–11949. doi:10.1364/OE.24.011932.
75. M. Mossaad, S. Hranilovic, L. Lampe, Visible light communications using OFDM and multiple LEDS, *IEEE Trans. Commun.* 63 (11) (2015) 4304–4313. doi:10.1109/TCOMM.2015.2469285.
76. H. Elgala, T. D. C. Little, Reverse polarity optical-OFDM (RPO-OFDM): dimming compatible OFDM for gigabit VLC links, *Opt. Express* 21 (20) (2013) 24288–24299. doi:10.1364/OE.21.024288.
77. The International Commission on Illumination (CIE), CIE 1931 Standard Colorimetric Observer, Retrieved from http://www.cie.co.at (Aug. 2008).
78. T. Fath, H. Haas, Performance comparison of MIMO techniques for optical wireless communications in indoor environments, *IEEE Trans. Commun.* 61 (2) (2013) 733–742. doi:10.1109/TCOMM.2012.120512.110578.
79. K. Ying, H. Qian, R. J. Baxley, S. Yao, Joint optimization of precoder and equalizer in MIMO VLC systems, *IEEE J. Sel. Areas Commun.* 33 (9) (2015) 1949–1958. doi:10.1109/JSAC.2015.2432515.
80. L. Zeng, D. O'Brien, H. Minh, G. Faulkner, K. Lee, D. Jung, Y. Oh, E. T. Won, High data rate Multiple Input Multiple Output (MIMO) optical wireless communications using white LED lighting, *IEEE J. Sel. Areas Commun.* 27 (9) (2009) 1654–1662. doi:10.1109/JSAC.2009.091215.
81. M. Safari, M. Uysal, Do we really need OSTBCS for free-space optical communication with direct detection?, *IEEE Trans. Wireless Commun.* 7 (11) (2008) 4445–4448. doi:10.1109/T-WC.2008.070637.
82. R. Mesleh, H. Haas, C. W. Ahn, S. Yun, Spatial modulation - a new low complexity spectral efficiency enhancing technique, in: *Proceedings of First International Conference on Communication and Networking in China*, Beijing, China, 2006, pp. 1–5. doi:10.1109/CHINACOM.2006.344658.
83. R. Mesleh, H. Elgala, H. Haas, Optical spatial modulation, *IEEE/OSA J. Opt. Commun. Netw.* 3 (3) (2011) 234–244. doi:10.1364/JOCN.3.000234.
84. T. Fath, M. D. Renzo, H. Haas, On the performance of Space Shift Keying for optical wireless communications, in: *IEEE GLOBECOM Workshops*, Miami, FL, USA, 2010, pp. 990–994. doi:10.1109/GLOCOMW.2010.5700474.
85. S. P. Alaka, T. L. Narasimhan, A. Chockalingam, Generalized spatial modulation in indoor wireless visible light communication, in: *IEEE Global Communication Conference*, San Diego, CA, USA, 2015, pp. 1–7. doi:10.1109/GLOCOM.2015.7416970.
86. T. Fath, C. Heller, H. Haas, Optical wireless transmitter employing discrete power level stepping, *J. Lightw. Technol.* 31 (11) (2013) 1734–1743. doi:10.1109/JLT.2013.2257984.
87. A. D. Griffiths, M. S. Islim, J. Herrnsdorf, J. J. D. McKendry, R. Henderson, H. Haas, E. Gu, M. D. Dawson, CMOS-integrated GaN LED array for discrete power level stepping in visible light communications, *Opt. Express* 25 (8) (2017) A338–A345. doi:10.1364/OE.25.00A338.
88. H. Ma, L. Lampe, S. Hranilovic, Coordinated broadcasting for multiuser indoor visible light communication systems, *IEEE Trans. Commun.* 63 (9) (2015) 3313–3324. doi:10.1109/TCOMM.2015.2452254.
89. Z. Chen, D. A. Basnayaka, H. Haas, Space division multiple access for optical attocell network using angle diversity transmitters, *J. Lightw. Technol.* 35 (11) (2017) 2118–2131. doi:10.1109/JLT.2017.2670367.

90. A. Nuwanpriya, S. W. Ho, C. S. Chen, Indoor MIMO Visible light communications: novel angle diversity receivers for mobile users, *IEEE J. Sel. Areas Commun.* 33 (9) (2015) 1780–1792. doi:10.1109/JSAC.2015.2432514.
91. L. Wei, H. Zhang, J. Song, Experimental demonstration of a cubic-receiver-based MIMO visible light communication system, *IEEE Photon. J.* 9 (1) (2017) 1–7. doi:10.1109/JPHOT.2016.2636752.
92. O. Narmanlioglu, R. C. Kizilirmak, T. Baykas, M. Uysal, Link adaptation for MIMO OFDM visible light communication systems, *IEEE Access* 5 (2017) 26006–26014. doi:10.1109/ACCESS.2017.2771333.
93. I. Tavakkolnia, C. Chen, R. Bian, H. Haas, Energy-efficient adaptive MIMO-VLC technique for indoor LiFi applications, 2018, p. accepted by International Conference on Telecommunications.
94. A. Goldsmith, *Wireless Communications*, Cambridge University Press, 2005.
95. J. G. Andrews, F. Baccelli, R. K. Ganti, A new tractable model for cellular coverage, in: *48th Annual Allerton Conference on Communication, Control, and Computing*, 2010, pp. 1204–1211. doi:10.1109/ALLERTON.2010.5707051.
96. A. A. Purwita, C. Chen, D. A. Basnayaka, H. Haas, Aggregate signal interference of downlink LiFi networks, in: *IEEE Global Communication Conference*, Singapore, 2017, pp. 1–6. doi:10.1109/GLOCOM.2017.8254782.
97. L. Yin, H. Haas, Coverage analysis of multiuser visible light communication networks, *IEEE Trans. Wireless Commun.* 17 (3) (2018) 1630–1643. doi:10.1109/TWC.2017.2782694.
98. H. S. Kim, D. R. Kim, S. H. Yang, Y. H. Son, S. K. Han, Mitigation of inter-cell interference utilizing carrier allocation in visible light communication system, *IEEE Commun. Letters* 16 (4) (2012) 526–529. doi:10.1109/LCOMM.2012.021612.112422.
99. K. Cui, J. Quan, Z. Xu, Performance of indoor optical femtocell by visible light communication, *Opt. Commun.* 298–299 (2013) 59–66. doi:10.1016/j.optcom.2013.02.018.
100. C. Chen, S. Videv, D. Tsonev, H. Haas, Fractional frequency reuse in DCO-OFDM-based optical attocell networks, *J. Lightw. Technol.* 33 (19) (2015) 3986–4000. doi:10.1109/JLT.2015.2458325.
101. D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, K. Sayana, Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges, *IEEE Commun. Mag.* 50 (2) (2012) 148–155. doi:10.1109/MCOM.2012.6146494.
102. C. Chen, H. Haas, Performance evaluation of downlink cooperative multipoint joint transmission in LiFi systems, in: *Proceedings of IEEE GLOBECOM Workshops*, Singapore, 2017, pp. 1–6. doi:10.1109/GLOCOMW.2017.8269151.
103. X. Li, F. Jin, R. Zhang, J. Wang, Z. Xu, L. Hanzo, Users first: user-centric cluster formation for interference-mitigation in visible-light networks, *IEEE Trans. Wireless Commun.* 15 (1) (2016) 39–53. doi:10.1109/TWC.2015.2466539.
104. Z. Chen, D. Tsonev, H. Haas, A novel double-source cell configuration for indoor optical attocell networks, in: *Proceedings of IEEE Global Communication Conference*, Austin, TX, USA, 2014, pp. 2125–2130. doi:10.1109/GLOCOM.2014.7037122.

32

Optical Data Storage

Byoung S. Ham

CONTENTS

32.1 von Neumann Architecture: CMOS Technologies 464
32.2 von Neumann Bottleneck 464
32.3 Non-von Neumann Architecture: Memcomputing 464
32.4 Non-volatile Photonic Memory 465
32.5 Photonic Memelements for Non-von Neumann Architecture 467
32.6 Other Method of Optical Memory 468
32.7 Associative Optical Memory for Pattern Recognition 468
References 470

Since the first laser invention in 1960 using a ruby crystal, laser technologies have been extensively developed for ultrafine lasers whose linewidth is far less than Hz, ultra-high power lasers whose power exceeds 100 kW, and ultrashort pulsed lasers whose pulse duration is as short as an attosecond. Since the first invention of semiconductor laser in 1962, development of semiconductor laser diodes has been intensively researched for a wide range of optical wavelengths ranging from blue to mid-IR and contributed to various optical applications such as optical communications, optical data storage, optical holography, optical sensors, and high-precision optical metrology. Especially the narrow linewidth lasers have been applied to broad range of scientific researches from basic optical spectroscopy in the 1970s to quantum sciences in the 2010s. For example, the use of ultrafine laser whose linewidth is sub-Hz plays an important role in optical data storage of a spectral hole-burning medium, atomic clocks for precision metrology such as GPS, and quantum information where a Rydberg atom is excited.

Over the last decades research on optical data storage has been intensively performed to support partially or to replace completely its electronic counterpart of CMOS technologies. Some have been successful, while others have failed. During the 1980s and 1990s, photon echoes were intensively investigated for all-optical data/image processing using dephasing and re-phasing of inhomogeneously broadened two-level atoms. In the late-1990s, photon echoes were challenged by DVD as well as holography benefited with versatile and rewritable storage capabilities. Due to the disadvantage of extremely low retrievable efficiency and relatively short storage time, the photon echo disappeared in the 2000s. In the early 2010s, however, the photon echo has revived upon the request of quantum information processing and played a key role in quantum memories. Because quantum technologies are completely different from classical ones, the quantum memory is excluded from the present chapter. However, new photon echo schemes are introduced to support the associative optical memory for parallel processing. The associative optical memories have not been successful yet due to technological deficiency but expected to contribute to new computing technologies with big data. The rewritable CD, DVD, and BD are briefly mentioned for both old technologies of storage only and the emerging technologies of an optical memelement for parallel processing.

Optical data storage has evolved its research direction from the old passive storage benefit to the future active memory that satisfies non-von Neumann architecture. Owing to the benefit of the Internet, we are living in the massive data era. The old subject of statistics has been highlighted as a major one in neuroscience, mathematics, and computer science to efficiently handle the decision problem, mimicking parallel processing in human brains composed of neurons and synapses. Current CMOS technologies, however, are not suitable for parallel processing but for serial processing satisfying von Neumann architecture based on the Turing machine. A Turing machine results in the present digital technologies of modern computation, where the processing unit is separated from the memory unit. Due to the necessity of big data processing, a new challenge of CMOS technologies has been confronted, resulting in, the so-called, von Neumann bottleneck. To solve the von Neumann bottleneck, parallel processing is inevitable. Thus, the paradigm shift from von-Neumann to non-von Neumann architecture in computing technologies seems to be obvious and timely suitable. Artificial intelligence (AI) based on a new paradigm may lead to a new technology in the future. In that scope, the optical data storage does not need to cling onto the old concept anymore and should challenge both functionalities of processing and storage in order to satisfy the non-von Neumann architecture, too.

32.1 von Neumann Architecture: CMOS Technologies

High-speed information processing has gained its interest due to potential applications of AI relying on big data [1–8]. For AI, the architecture of a deep neural network (DNN) has been adapted to the data processing, where parallel processing plays a key role [9]. However, based on the von Neumann architecture applied to the modern semiconductor CMOS technology, the processing in the central processing unit (CPU) is not parallel but rather serial with respect to that in the memory [1]. Moreover, the data memory unit in the von Neumann architecture is spatially separated via electrical transmission lines. According to the Moore's law, the CPU speed (or number of transistors per chip) has doubled every 18 months over decades as shown in Figure 32.1 [10]. Due to lithography and current leakage (heating) problems, however, such VLSI evolution has been halted, and thus the evolution of Moor's law determining the CPU speed has been stopped at ~4 GHz at last [11].

32.2 von Neumann Bottleneck

In the von Neumann architecture, the CPU unit is distinctly separated from the memory unit, where the data transfer between these two units becomes a major hurdle if the amount of data increases [1–8,12]. At a glance, this seems to be true because of the limited bandwidth of a few GHz in the copper-based electronic transmission line. However, the main hurdle originates in the architecture of the Turing machine unsuitable for solving non-deterministic polynomial (NP) problems in AI. Thus, the data bottleneck between CPU and memory units has become a main issue in the current digital computing devices based on the von Neumann architecture, resulting in, the so-called, von Neumann bottleneck [5]. Recently, parallel processing has gained its importance for future technologies especially for big data processing, where neuromorphic computation is the first step forward [4], and quantum computing may soon follow [11]. As a result, the graphic processing unit has already emerged and dominated big data processing in the conventional von Neumann architecture. The cease of Moor's law is of course due to the technological limit on one hand, but on the other hand, it is inevitable to solve the von Neumann bottleneck with big data. This may explain how multi-core CPU has been evolved in the COMS technologies (see Figure 32.1).

32.3 Non-von Neumann Architecture: Memcomputing

To solve the von Neumann bottleneck in current COMS technologies, the idea of memcomputing has emerged a decade ago, which allows for concurrent processing while storing the result

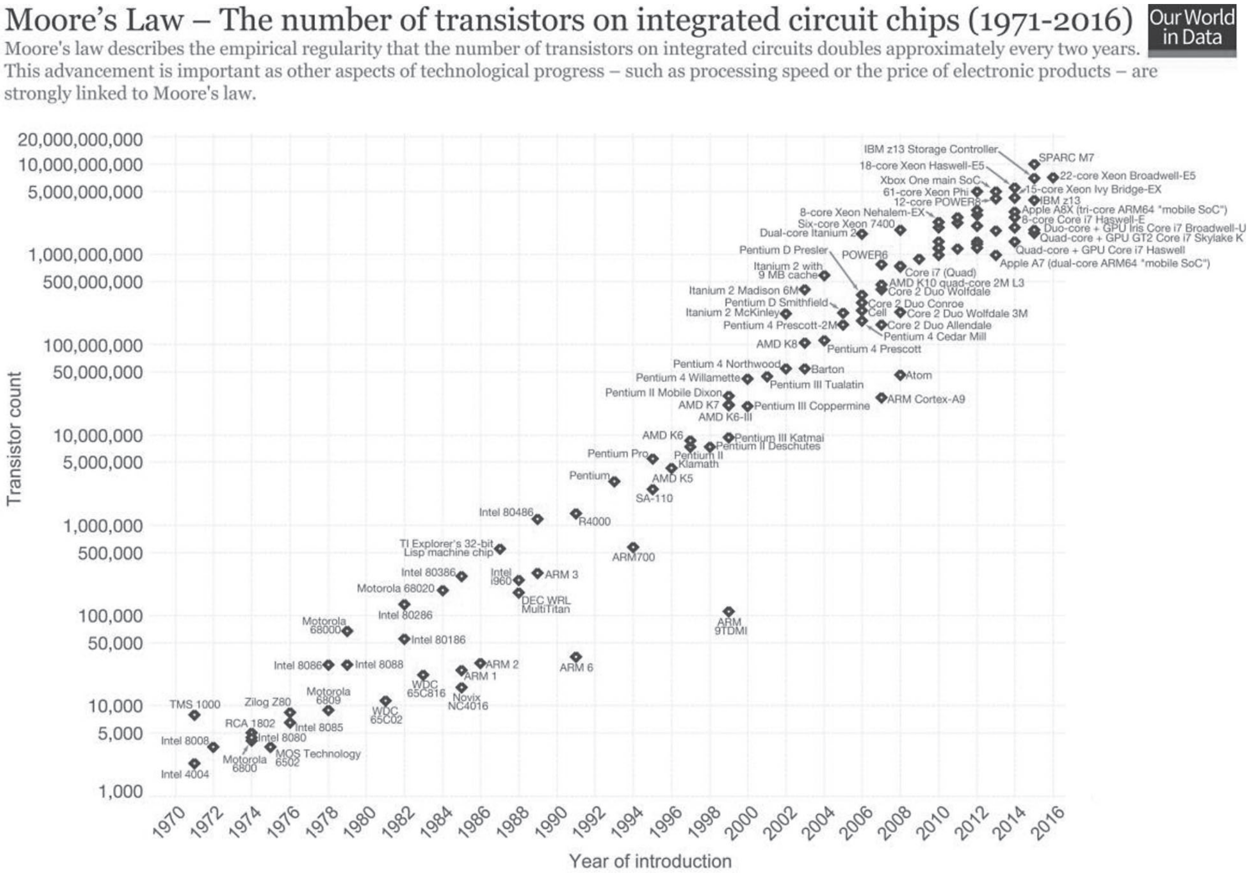


FIGURE 32.1 Moore's law. https://ourworldindata.org/

in the same unit [1–3,13,14]. Although memcomputing is still in the very early stages of its technological evolution [1], it represents a paradigm shift in computation from von Neumann to non-von Neumann architecture. Memcomputing also implies non-Markovian regime, where the past history is the source of the future outcome [15]. In other words, the processing result is kept at the memelement unit and used as a source for the next processing. The main hardware of the memcomputing is called a memelement. One of the memelements is memristor, where both logic operation and storage occur simultaneously in the same physical unit. The target of memcomputing is of course for parallel processing to mimic a neural network of a human brain function composed of neurons and synapses.

Before getting into the main topic of the optical data storage, it may be helpful to the general audience to introduce the background of the parallel computing approach beyond the von Neumann architecture. An artificial neural network mimics a human brain, where neurons play a role of data processing, while the synapses play a role of storage. Here, the number of synapses far exceeds the number of neurons by a factor of 10^4 in the human brain [1]. In the architecture of neuromorphic computing, a conditional probability acts a major role like in the synaptic connections between the neuron's layers [4]. The conditional probability implies analog rather than digital means.

Recently various neuromorphic hardware chips have been developed to mimic the brain functions of parallel processing to overcome the von Neumann bottleneck with the help of neuroscience [3–7]. So far most research on neuromorphic computing has been focused on software engineering to emulate brain functionality such as image processing and voice-recognition targeting for deep learning. For the implementation of memcomputing, one of the most important requirements is a non-volatile memory. In modern CMOS technologies based on flip-flop and latch circuits, the fundamental problem to neuromorphic computing is the volatile memories such as dynamic random-access memory (DRAM) and static random-access memory (SRAM). To fit it into the synaptic applications of memcomputing, several types of memories have been developed, among which are resistive switching memory (RRAM), phase change memory (PCM), and magnetoresistive memory (MRAM). Although these non-volatile memory technologies are in the early stage, several neuromorphic chips such as IBM's True North and Qualcomm's NPU have already been developed.

Since the first suggestion of memristor in the 1970s [13], modified CMOS technologies have been shown for demonstrating synaptic behaviours [1–8,14]. Deep learning is a type of machine learning based on DNN, where hidden layers play a major role in terms of learning process [9]. Current deep learning technology is mostly based on software engineering in the von Neumann architecture. The hidden layers offer non-linear classification boundaries, where they are composed of binary neurons and analog synapses. The most often used type of deep learning is a sigmoid neuron which is a binary that determines the output depending on its weight. The major problem in deep learning is the requirement of a substantial amount of computing resources, and this requirement becomes more severe as the hidden layers get thicker. The best solution is to develop the parallel computation hardware. To implement neuromorphic hardware, at least two major components are required: one is an induction protocol and the other is a synaptic weight memory block. The major challenge is to get these two different functional blocks into a single two-terminal device. Another requirement of the memory is to have a long-term storage capability, i.e. a non-volatile functionality.

Memcomputing stands for non-von Neumann architecture, where the main hardware is a memelement like memristor [1,14]. The electrical/optical properties of the memelement of a two-terminal electronic/optical device must be kept in its record history. The general definition of memelements or memory circuit elements is the time-dependence in its output. Computation is thus achieved through the time evolution of the circuit when it is connected to external resources. The collective dynamics of the memelements results in energy-efficient computing power via massively parallel processing for the hidden layers of synaptic elements.

There are six most important criteria for memcomputing [5]. First, the memelement should perform both functions of processing and storage simultaneously in an analog form. Thus, memcomputing reduces the computing resources and bypasses the von Neumann bottleneck between computing elements and memory blocks. Second, the storage time in the memelement should be much longer than the processing time. To satisfy this condition, a non-volatile storage device is essential. Fast read/write time is also important. In current CMOS technologies, sub-ns switching time is confirmed with more than 10^5 cycled operations whose retention is several years. For example, Ta_2O_{5-x}/TaO_{2-x} bilayer structures sustain up to 10^{12} write cycles. Third, initialization must be provided for the memelement before the computation begins. Fourth, the memelement must satisfy collective dynamics, where time evolution on a memelement state depends on the states of all other elements. This means that the memdevice should have strong memory contents. Fifth, the final state can be read out reliably without modifying other elements' status. Sixth, the memelement should be resistant against environmental noise. This means that the memcomputing device needs to be insensitive to small variations in individual components or perturbations. The main purpose of memcomputing is to efficiently solve NP hard problems defined by the computational complexity theory such as the maze problem [16] and the shortest path problem [17], where massive parallel processing is an essential requirement.

32.4 Non-volatile Photonic Memory

Phase change materials (PCMs) are widely used for both optical and electrical data storages especially for rewritable optical/electrical data storage devices such as rewritable DVDs and phase change random access memory (RAM) devices [18–20]. The major function of PCM is a reversible process between amorphous and crystalline phases upon heating by a laser pulse or electric current. The typical types of PCMs are Ge-Sb-Te (GST) and AgInSbTe (AIST). The hysteresis of PCM is reliable via given two distinct amorphization and crystallization phases determined by the meting temperature. The PCM materials also satisfy the non-volatile requirement of memelements.

The major function of non-volatile photonic memory is the capability of multi-level RAM that satisfies memelements. The PCM technique uses refractive index difference between two distinct phases of amorphous and crystalline states [19–21]. The initial state can be either amorphous or crystalline depending on its purpose. If PCM is initialized into the amorphous state, applying a certain amount of heat changes it to the crystalline state, whose refractive index is higher or absorptive, resulting in less transmission or weak signal detection of the probe light. Usually the optical distinction ratio between the amorphous and crystalline phases is about 20% in a GST-PCM material [21].

From the competition between magneto-optical and PCM in rewritable optical storage media, PCM survives as the winner in the rewritable product due to the discovery of alloys of Ge, Sb, and Te (see Figure 32.2). The crucial properties of phase change alloys are as follows [22].

1. High-speed phase transition
2. Long thermal stability of amorphous state
3. Large optical change between the two states for rewritable storage
4. Large cycle number of reversible transitions

Starting from the early discovery of PCM based on Te such as $Te_{85}Ge_{15}$ doped with elements such as Sb, S, and P whose crystallization is of the order of microseconds, the first fast recrystallization and good optical contrast were found in GeTe and GeTe-SbTe, where this alloy has been widely used for rewritable DVDs (DVD-RAM) in the 1990s and Blue ray disks (BD) in the 2000s. The second material family of doped SbTe alloys was found later where $Ag_5In5Sb_{60}Te_{30}$ (AIST) became the major one in rewritable optical storage media such as DVD-RW in the 1990s.

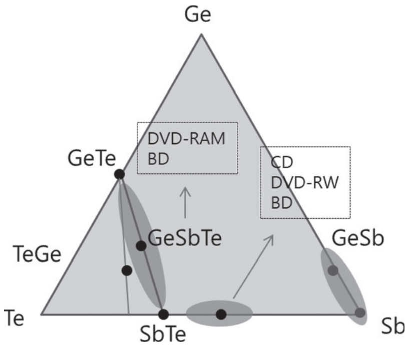


FIGURE 32.2 GST PCM material.

The amorphous phase has high resistance (or low refractive index or high optical transmission), while the crystalline phase has low resistance (or high refractive index or low optical transmission) [19–22]. When applying heating on the amorphous phase above the meting temperature, the amorphous phase turns into the crystalline phase. Higher heating applied to the crystalline state induces local melting, resulting in amorphous localization on a rapid process (see Figure 32.3): recrystallization. Thus, the reversible process between two phases is possible by adjusting the heating process.

For the purpose of memelements, the fast switching time between two distinct phases is one of the most important parameters, which is determined by the recrystallization process below the melting temperature, but above the glass-transition temperature. This means that there exists a hysteresis curve representing PCM plasticity required for the non-volatile memory.

Figure 32.4 shows various conventional optical data storage media of rewritable compact disk (CD-RW), rewritable digital versatile disk (DVD-RAM), and Blue-ray rewritable disk (BD-RE) in GST-PCM materials. These conventional media have been successful for non-volatile optical storages like a hard disk in an electronic version. As shown in Figure 32.4, the recording density has been increased by a factor of 7 from

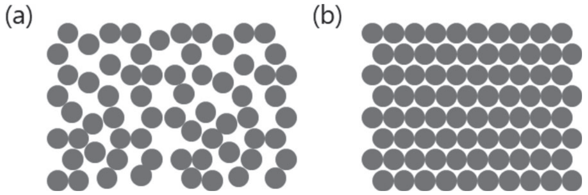


FIGURE 32.3 (a) Amorphous phase: low reflectivity (high transmission). (b) Crystalline phase: high reflectivity (low transmission).

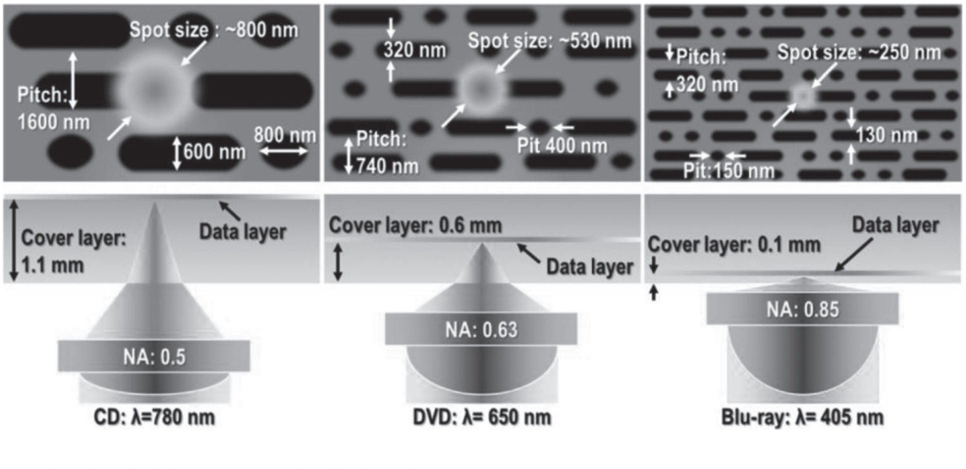


FIGURE 32.4 Conventional optical data storage technologies using PCM.

CD-RW to DVD-RAM in the 1990s. From DVD-RAM to BD-RE the density increases by a factor of 5 in the early 2000s. The rewriteability of DVD-RAM is about 10 000 times, and the data can be retained for about 30 years. Here, the density is only proportional to the laser spot size on the surface of the media, where the minimum pit size should be dependent upon its wavelength. The rewriting speed is also dependent on the recrystallization process. Although these non-volatile optical data storage technologies are based on the phase change in PCM, its mechanism is conceptually old and losing competence with respect to the electronic counterparts. As a result, Panasonic has announced that it will end the production of DVD-RAM by mid-2019. Recently, the GST-PCM has been applied to non-volatile electronic devices such as flash memories competing with the volatile memories of DRAM and SRAM with faster switching time in the phase transition.

32.5 Photonic Memelements for Non-von Neumann Architecture

Figure 32.5 shows a schematic on how the phase-change control is performed in PCM for both electronic/optical versions [18]. Because the physics of the phase change in PCM should rely on amorphization and crystallization as done in rewritable DC, DVD, and BD, the phase change control is actually the same as in all cases via heating with an electrical or optical pulse. For the function of heating applied to a PCM device, the ‘SET’ pulse is used for recrystallization from the amorphous phase by simply slowly heating to just above the crystallization temperature for longer period. The short ‘RESET’ pulse is to turn the crystalline phase into the amorphous phase by heating over the melting temperature (~600°C). This ‘RESET’ represents the ‘WRITE’ process, resulting in the logical bit ‘1’. The read operation is performed by measurement of the device resistance (transmission) at low voltage (weak light field), where the low voltage (weak light) is to protect the device from any perturbation to the given phase state. To be worked as a memelement, the most important function of this type of non-volatile PCM device is to store more than 1 bit per cell, resulting in a multi-level memory device with analog resistance (refractive index) states. The amorphous state has a few orders of magnitude higher electrical resistivity compared with that of the crystalline phase.

For the photonic non-volatile memory the amorphous phase has less optical absorption (or high optical transmission) compared with the crystalline phase of PCM. Thus, the ‘WRITE’ process is performed by the ‘RESET’ resulting in more transmission to the weak probe field, corresponding to the case of less electrical current as shown in Figure 32.5. Figure 32.6 shows a schematic of two distinctive PCM phases reversibly operated by ‘RESET’ and ‘SET’ pulses [18]. Figure 7 shows a schematic of the photonic PCM corresponding to Figure 32.5 [21]. The switching mechanism is in the phase change between amorphous and crystalline states via optical pulse activated for local heating.

Figure 8 shows about 20% transmission distinction between two phases of GST-PCM and reversible operations at GHz writing speed for five different multi-level storages. The switching pulse energy is as low as 13 pJ. Here, the writing (amorphization) speed is a more stringent requirement as it determines the information storage speed. Over the last decade, several PCM materials have been investigated intensively to understand the phase mechanism at a faster writing speed.

In 2017, the photonic non-volatile memory using PCM has been developed to scalable memelements for memcomputing, where the memelement is composed of individual PCM cells located at grid points in a two-dimensional lattice structure. All PCM cells are connected by optical waveguides crossing them [23]. As mentioned above in Figures 32.5–32.7, a multi-level memory function of each PCM cell using ‘RESET’ and ‘SET’ optical pulses has been performed for photonic processing of arithmetic calculations in consecutive two PCM cells, mimicking the ancient abacus. In this case, a carryover in the abacus was also demonstrated in the second PCM cell, where the full crystallization resulting from the addition process by multi-steps of the ‘RESET’ in the first PCM cell triggers storage in the second cell.

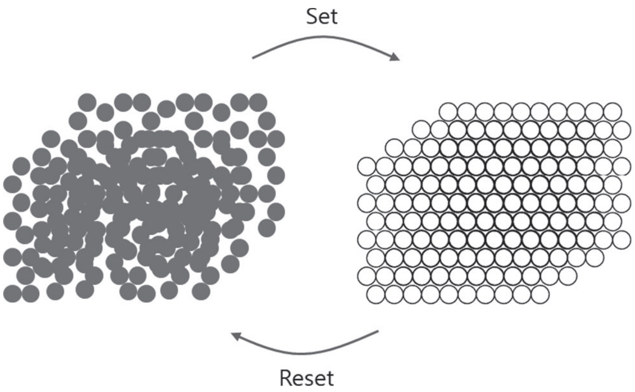


FIGURE 32.6 Switching reversibility in PCM by using set and reset pulses.

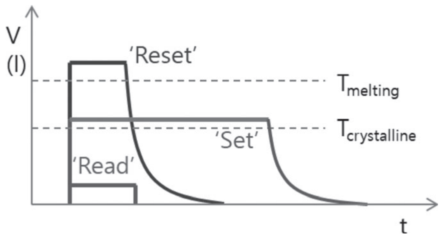


FIGURE 32.5 Reversible programming in a PCM device.

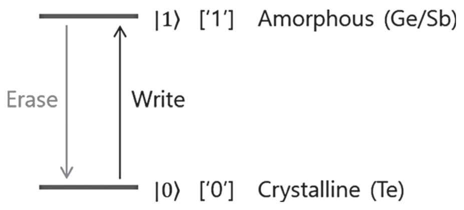


FIGURE 32.7 A schematic of photonic PCM between two distinct phases.

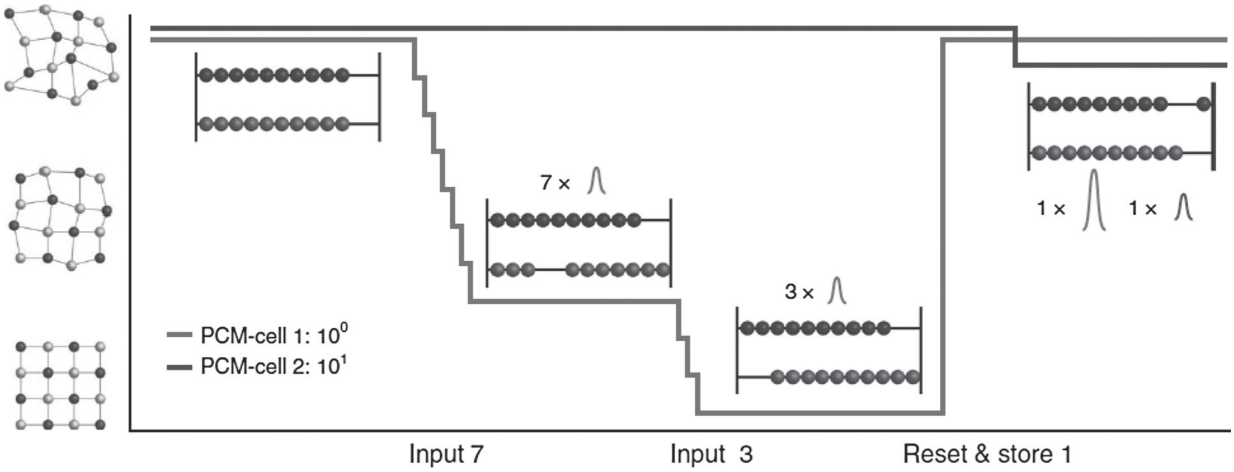


FIGURE 32.8 A schematic of photonic memelement using two different PCM cells connected by optical waveguides [23].

In Figure 32.8, the basic function in a PCM-based memelement has been experimentally demonstrated at a GHz speed for arithmetic calculations via analog data processing [23]. The initial state is the amorphous phase of a GST-PCM material. The duration of each 'RESET' pulse is ~ps whose energy is 12 pJ at a telecommunication C-band wavelength of 1530–1565 nm. Unlike Figure 32.7, the amorphous phase is considered as the logical bit '0' in Figure 32.8. Thus, the initialization of both PCM cells in Figure 32.8 is for amorphization representing the logical bit '0'. As a result, the writing process is the recrystallization. The number of bit levels is designed to be 10, representing a 10-base arithmetic calculator. The 10-red-bead rod (second row in each bead-rod set) represents the first 10-level PCM cell, while the 10-blue-bead rod (first row in each bead-rod set) represents the second 10-level PCM cell. Initially, all digits are zeros, where all beads stay in the left side. The number 7 is activated by seven consecutive optical pulses, resulting in seven steps down in the transmission profile: see the red-line steps and the red bead's relocation to the right side in that rod. Then, the number 3 is added by three more consecutive light pulses, resulting in full recrystallization: see all 10 red beads are moved to the right side of the rod. Once the PCM cell reaches the full recrystallization phase, a 'RESET' pulse fires, changing the PCM into an amorphous phase, and then the extra energy of the 'RESET' pulse is used in the next recrystallization step in the second PCM cell represented by the blue beads. Thus, the arithmetic addition of '7+3=10' is achieved experimentally in the two-connected PCM cells by an optical waveguide. The read-out pulse is very weak not to perturb the PCM cells. This arithmetic process demonstrates the breakdown of the well-known von Neumann bottleneck and shows the potential of photonic non-volatile memelement for memcomputing [24].

32.6 Other Method of Optical Memory

Besides the PCM materials for memelements, a photonic crystal cavity has been demonstrated as a RAM in fast all-optical data processing, where RAM is volatile necessary in the von Neumann architecture [25]. In this type of optical RAM, the quality factor of a nano-cavity in the photonic crystal determines the storage time. The multiple channel RAMs can be fabricated for different wavelength photonic crystal structures in the same chip. In Ref. [25], NTT demonstrated large-scale integration of wavelength-addressable all-optical memories on a photonic crystal chip made of either III-V semiconductor compounds or silicon. The number of demonstrated memory cells is up to 105, and the writing time is down to sub-ns.

The RIKEN group of Japan has demonstrated an all-optical volatile memory in a carbon nanotube (CNT), where CNT can be integrated with silicon photonics [26]. The group demonstrated reversible and reproducible memory operations in a micron-size CNT with optical bistability at sub-mW excitation power. The writing speed is in a sub-ms range for a μm-long CNT device.

32.7 Associative Optical Memory for Pattern Recognition

In the area of all-optical information processing, an optical associative memory (OAM) is a key element for all-optical pattern recognition, and pattern recognition is the key subject of DNN in AI. During the 1980s and 1990s, photon echoes and holography were widely investigated for potential applications of OAM, but potentially failed due to extremely low retrieval efficiency for feedback processing, where the feedback plays a crucial role in the associative memory [27–34]. In the early 2000s, a backward photon echo scheme was suggested for a near-perfect retrieval efficiency in an atomic vapour medium [35]. This near-perfect photon-echo scheme was originally proposed for quantum memories. In the late 2010s, the backward photon echo scheme was modified, applied to a solid ensemble, and demonstrated the proof-of-principle of near-perfect photon echoes with an increment of 15 times in the retrieval efficiency [36]. Moreover, ultra-long storage time of a quantum memory was suggested in the late 2000s and demonstrated for spin population decay time, which is the longest storage time in the dynamic optical data storage [37]. Thus, the associative memory is ready for all-optical pattern recognition

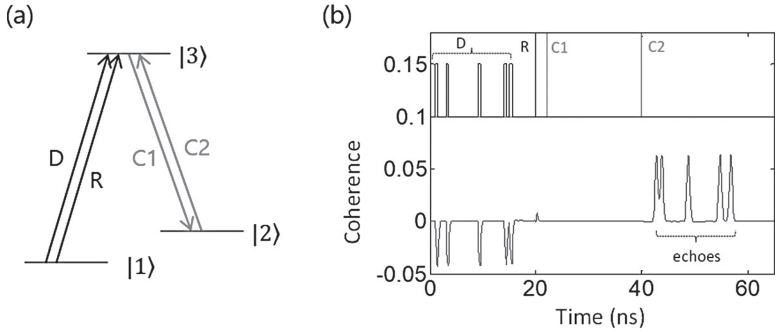


FIGURE 32.9 Quantum coherence control applied to backward photon echoes in a three-level system. (a) Energy-level diagram of a solid medium. (b) Numerical calculations of (a). The pulse C2 only is backward.

in principle owing to the ultra-long storage time as well as near-perfect retrieval efficiency in the modified photon echoes.

The inset of Figure 32.9b represents the related pulse sequence for Figure 32.9a, where the numerical calculations by solving density matrix equations are shown in the bottom pannel of Figure 32.9b. The D and R indicate data and re-phasing pulses, respectively, where the re-phasing pulse has a 2π pulse area and used for swapping populations between states |1> and |3> [36–39]. The data D is composed of five weak consecutive optical pulses, where the photon echo is appeared in a reverse order as shown. The pulses C1 and C2 are for the quantum coherence control, whose function is the coherence transfer between the optical and spin transitions, i.e. ρ_{13} and ρ_{12}. This kind of quantum coherence control has been intensively studied in a three-level optical system for ultra-slow light-based quantum memories [40]. The theoretical calculations of the coherence conversion between optical and spin transitions have also been studied in a rare-earth-doped solid ensemble. This study was completed using numerical calculations to solve the nine density matrix equations interacted with the light pulses in time domain. The resulting wave propagation direction and wavelength of the photon echo are determined by the following non-degenerate four-wave mixing processes:

$$\boldsymbol{k}_{\mathrm{e}} = -\boldsymbol{k}_{\mathrm{C1}} + \boldsymbol{k}_{\mathrm{C2}} + \boldsymbol{k}_{\mathrm{D}} \sim -\boldsymbol{k}_{\mathrm{D}}, \tag{32.1}$$

$$\omega_{\mathrm{e}} = -\omega_{\mathrm{C1}} + \omega_{\mathrm{C2}} + \omega_{\mathrm{D}} = \omega_{\mathrm{D}}, \tag{32.2}$$

where $\mathbf{k}_{\mathrm{e}}$ and ω_{e} are the propagation and wavelength of the echo signal, respectively. As shown in equation (32.1), the propagation direction of the echo is not exactly opposite of the data. However, as shown in the preliminary experiment in a rare-earth doped ensemble, the overlap among the pulses has been achieved at ~ 90%, and the enhanced echo efficiency has been observed [36]. Here, it should be noted that a rare-earth doped solid ensemble is spin inhomogeneous. Thus, conventional quantum coherence control does not apply. To solve this spin inhomogeneity problem, Raman re-phasing is added to the conventional quantum coherence control to overcome the shortened spin dephasing [38,39]. The optical Raman pulses can also be replaced by a direct microwave pulse for spin re-phasing [41]. If it is for a quantum memory, then the second control (readout) pulse must be resonant to the ground state |1> rather than to state |2> to avoid of spontaneous emission noise: $\omega_{C2} = \omega_D$; $\mathbf{k}_{\mathrm{C2}} + \mathbf{k}_{\mathrm{D}} = 0$ [39]. In the applications of general optical memory, however, the spin state choice does not matter unless the phase-matching condition is not satisfied. From equation (32.1), the following equation can be satisfied if the second control pulse C2 counter-propagates with respect to that of the data pulse D:

$$\boldsymbol{k}_{\mathrm{e}} = -\boldsymbol{k}_{\mathrm{C1}} + \boldsymbol{k}_{\mathrm{C2}} + \boldsymbol{k}_{\mathrm{D}} = -\boldsymbol{k}_{\mathrm{C1}}. \tag{32.3}$$

Figure 32.10 shows theoretical calculations for both backward and forward photon echoes. As shown, the retrieval efficiency of the backward photon echo becomes near perfect as the optical depth increases, whilst it diminishes exponentially in the forward case as already known, resulting in less than 1% echo efficiency in conventional photon echoes.

The associative memory is defined as the ability to learn or to remember the relationship between unrelated or partial items [29–34]. This may imply remembering the name of someone or the aroma of a particular perfume. This type of memory deals specifically with the relationship between these different objects or concepts. In a normal associative memory, the recall of pairs of unrelated items is tested, e.g. for face-name relations. A regular memory is for a storage location of a particular address. An example is a house on a street with a corresponding street address. Associative memory is for a set of storage locations that are accessed through the contents rather than the address. So, the associative memory must be content addressable, where it opens for a parallel access. Thus,

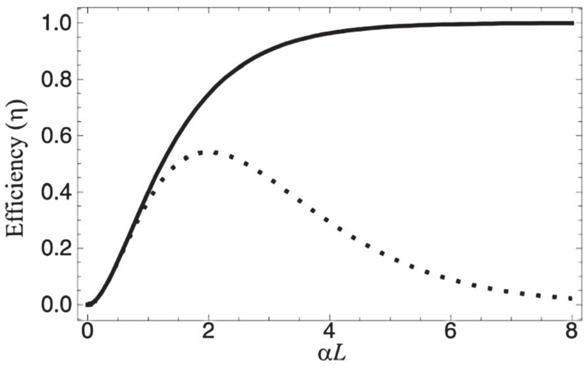


FIGURE 32.10 Retrieval efficiency of photon echoes. Solid curve: backward scheme. Dotted curve: forward scheme. The αL indicates optical depth.

the associative memory conflicts with von Neumann architecture and represents the heart of non-von Neumann processing. The physics of associative memory is in the autocorrelation/convolution functions [29,34]. Thus, it can be applied to fault-tolerant content addressable memories, too. In AI, Hopfield network is closely related with the associative memory, where it is also called Hebbian learning [42]. Thus, optical pattern recognition is inherently based on parallelism with feedback process such as back propagation in DNN [3,9].

In the late 1980s, holographic storage has been suggested for all-OAM using correlation and convolution as shown in Figure 32.11 [31]. Several different images are stored in the hologram upon reference teams with different angles. To read out the stored image, suppose that an incomplete image $\hat{a}_i$ (W) is sent to the hologram. Then, the correlated but distorted reference beam $\hat{b}_i$ comes out, where the reference beam $\hat{b}_i$ is supposed to feed into the phase conjugate material (PCJM) under a certain additional non-linear optical processes, resulting in a phase conjugate beam $\hat{b}_i^*$ retrieved from the PCJM. The retrieved phase conjugate $\hat{b}_i^*$ is now fed back into the hologram, and then the full image of a_i is retrieved at last as shown in Figure 32.11. This image is the associated image through the associative memory process. If this image is not clear enough, the process mentioned above is repeated until the original image is retrieved. Here, the major problem is efficiency, where the intensity of $\hat{b}_i^*$ should be comparable to $\hat{b}_i$, which has become a major challenge in quantum non-linear optics so far.

The followings are the expressions of the related non-linear optical process for the associative memory in Figure 32.11, where $\hat{A}_i\bar{A}_i$ represents the correlation in equation (32.4) [30]:

$$\hat{A}|A+B|^2 = \hat{A}\left(|A|^2+|B|^2\right) + \hat{A}A\bar{B} + \hat{A}\bar{A}B, \qquad (32.4)$$

$$\hat{A}\bar{A}B|A+B|^2 = \hat{A}\bar{A}B\left(|A|^2+|B|^2\right) + \hat{A}\bar{A}BA\bar{B} + \hat{A}\bar{A}B\bar{A}B$$

$$= \hat{A}\bar{A}B\left(|A|^2+|B|^2\right) + \hat{A}\bar{A}|B|^2A + \hat{A}\bar{A}^2B^2. \qquad (32.5)$$

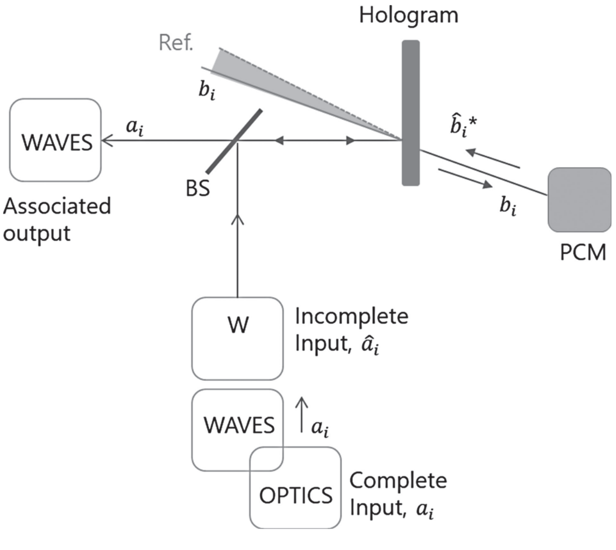


FIGURE 32.11 A schematic of all-OAM

Equation (32.4) represents the read-out process to PCJM, while equation (32.5) represents the feedback process. The associative memory is denoted by the red A in the second-to-last item of equation (32.5). This theory cannot be applied until retrieval efficiency in the PCM is good enough.

Photon echoes are similar to holography in terms of the four-wave non-linear optical processes [27–34]. The original version of photon echoes that flourished in the 1980s and 1990s is for a two-level solid bulk material, where the echo retrieval efficiency is normally less than 1% of the stored data due to the echo reabsorption by the medium in the same propagation direction [27–29]. To overcome this ultra-low echo efficiency, the backward photon echo scheme was proposed and demonstrated in a three-level system with quantum coherence control [35,36]. Although the experimental proof using optical images has not been performed yet, the OAM should become an active area for the future non-von Neumann architecture or memcomputing as discussed earlier.

REFERENCES

1. Jeong D. S. et al. 2016 Memristors for energy-efficient new computing paradigms *Adv. Electron. Mater.* **2**, 1600090.
2. Burr G. W. et al. 2015 Experimental demonstration and tolerancing of a large-sacle neural network using phase-change memory as the synaptic weight element *IEEE Trans. Elec. Dev.* **62**, 3498–3507
3. Narayanan P. et al. 2017 Toward on-chip acceleration of the backpropagation algorithm using nonvolatile memory 2017 *IBM J. Res. & Dev.* **61**, 11.
4. Pershin Y. V., Di Ventra M. 2011 Neuromorphic, digital, and quantum computation with memory circuit elements *Proc. IEEE* **100**, 2071–2080
5. Di Ventra M., Pershin Y. V. 2013 The parallel approach *Nat. Phys.* **9**, 200–202.
6. Krestinskaya O., Salama K. N., James A. P. 2018 Learning in memristive neural network architectures using analog backpropagation circuits *IEEE Trans. Circuits & Sys.* **66**, 719–732.
7. Pedretti G. et al. 2017 Memristive neural network for on-line learning and tracking with brain-inspired spike timting dependent plasticity *Sci. Rep.* **7**, 5288.
8. Le Gallo M. et al. 2017 Mixed-precision in-memory computing *Nat. Electron.* **1**, 246–253.
9. LeCun Y., Bengio Y., Hinton G. 2015 Deep learning *Nature* **521**, 436–444.
10. Moore G. M. 1965 Cramming more components onto integrated circuits *Electronics* **38**, 8.
11. Nielsen M. A., Chuang I. L. 2010 *Quantum Computation and Quantum Information* (Cambridge University Press, New York).
12. Merolla P. A. et al. 2014 A million spiking-neuron integrated circuit with a scalable communication network and interface *Science* **345**, 668–673.
13. Chua L. O. 1971 Memristor-the missing circuit element *IEEE Trans. Circuit Theory* **CT-18**, 507–519.
14. Strukov D. B., Sinder G. S., Stewart D. R., Williams R. S. 2008 The missing memristor found *Nature* **453**, 80–83.
15. Takacs L. 1966 *Stochastic Process: Problems and Solutions* (London, Champman and Hall).

16. Pershin Y. V., Di Ventra M. 2011 Solving mazes with memristors: A massively parallel approach *Phys. Rev. A* **84**, 046703.
17. Pershin Y. V., Di Ventra M. 2013 Self-organization and solution of shortest-path optimization problems with memristive networks *Phys. Rev. A* **88**, 0133305.
18. Burr G. W. et al. 2016 Recent progress in phase-change memory technology *IEEE J. Emerg. Selec. Top. Circuits & Sys.* **6**, 146–162.
19. Wuttig M., Bhaskaran H., Taubner T. 2017 Phase-change materials for non-volatile photonic applications *Nat. Photon.* **11**, 465–476.
20. Shportko K. et al. 2008 Resonant bonding in crystalline phase-change materials *Nat. Mater.* **7**, 653–658.
21. Rios C. et al. 2015 Integrated all-photonic non-volatile multi-level memory *Nat. Photon.* **9**, 725–732.
22. Wuttig M., Yamada N. 2007 Phase-change materials for rewriteable data storage *Nat. Mater.* **6**, 824–832
23. Feldmann J. et al. 2017 Calculating with light using a chip-scale all-optical abacus *Nat. Commun.* **8**, 1256.
24. Traversa L. F., Di Ventra M. 2015 Universal memcomputing machines *IEEE Trans. Neural Net. & Learning Sys.* **26**, 2702–2715.
25. Kuramochi E. et al. 2014 Large-scale integration of wavelength-addressable all-optical memories on a photonic crystal chip *Nat. Photon.* **8**, 474–481.
26. Uda T., Ishii A., Kato Y. K. 2017 Single carbon nanotubes as ultrasmall all-optical memories *ACS Photon.* **5**, 559–565.
27. Carlson N. W., Babbitt W. R., Mossberg T. W. 1983 Storage and phase conjugation of light using stimulated photon echoes *Opt. Lett.* **8**, 623–625.
28. Xu E. Y., Kroll S., Huestis D. L., Kachru R., Kim M. K. 1990 Nanosecond image processing using stimulated photon echoes *Opt. Lett.* **15**, 562–564.
29. Belov M. N. and Manykin E. A. 1991 Optical associative memory based on an echo correlator *Opt. Lett.* **16**, 327–329.
30. Soffer B. H., Dunning G. J., Owechko Y., Marom E. 1986 Associative holographic memory with feedback suing phase conjugate mirrors *Opt. Lett.* **11**, 118–120.
31. Dunning G. J., Marom E., Owechko Y., Soffer B. H. 1987 All-optical associative memory with shift invariance and multiple-image recall *Opt. Lett.* **12**, 346–348.
32. Rizvi A. A. and Zubairy M. S. 1994 Implementation of associative memory using grating structure *Appl. Opt.* **33**, 3642–3646.
33. Ishikawa M., Mukohzaka N., Toyoda H., Suzuki Y. 1990 Experimental studies on learning capabilities of optical associative memory *Appl. Opt.* **29**, 289–295
34. Owechko Y., Dunning G. J., Marom E., Soffer B. H. 1987 Holographic associative memory with nonlinearities in the correlation domain *Appl. Opt.* **26**, 1900–1910.
35. Moiseev S. A., Kroll S. 2001 Complete reconstruction of the quantum state of a single-photon wave packet absorbed by a Doppler broadened transition. *Phys. Rev. Lett.* **87**, 173601.
36. Hahn J. and Ham B. S. 2011 Rephasing halted photon echoes using controlled optical deshelving *New J Phys.* **13**, 093001.
37. Ham B. S. 2012 Atom phase-locked coherence conversion using optical locking for ultralong photon storage beyond the spin T2 constraint *New J Phys.* **14**, 013003.
38. Turukhin A. V. et al. 2002 Observation of ultraslow and stored light pulses in a solid *Phys. Rev. Lett.* **88**, 023602.
39. Ham B. S. 2018 A wavelength convertible quantum memory: Controlled echo *Sci. Rep.* **8**, 10675.
40. Lukin M. D. 2003 Colloquium: Trapping and manipulating photon states in atomic ensembles *Rev. Mod. Phys.* **75**, 457–472.
41. Laplane C., Jobez P., Etesse J., Gisin N., Afzelius M. 2017 Multimode and long-lived quantum correlations between photons and spins in a crystal *Phys. Rev. Lett.* **118**, 210501.
42. Keysers C, Gazzola V 2014 Hebbian learning and predictive mirror neurons for actions, sensations and emotions *Phil. Trans. R. Soc. B* **369**, 20130175.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Index

Note: **Bold** page numbers refer to **tables** and *italic* page numbers refer to *figures*.

Abbe errors 16
ablative photodecomposition *see* photoablation
absolute gravimetry
  cold atom gravimeters 176–8
  falling corner cube gravimeters 174–6, *175*
absorption coefficient 207, *207*
absorption spectrum 323
accidents, laser (case studies) 346–8
accuracy and quality control 193
acousto-optic frequency modulators (AOM) 6, 88
advanced LIGO (aLIGO) 18
advances in laser satellite communications
  applications 393
  atmospheric channel 391–2
  beam pointing and stabilization 387
  flight laser transmitter 391
  free-space optical communications signal flow 386, *387*
  ground station 392–3
  laser beam pointing 386–7
  lasercom benefits 384, *384*
  lasercom challenges 384–5
  link budget **385,** 385–6
  modulation and coding 390–1
  rugged opto-mechanical and thermo-mechanical optics *388,* 388–9, *389, 390*
  signal detection 389–90
  spatial acquisition 387
  technology maturity 385, **385**
  tracking 387
age-related macular degeneration (AMD) 232
Ajiasi 187
Albrecht, H.-E. 35, 39
4 +1 algorithm 66, *66*
4 +4 algorithm 65–6, *66*
Allan variance 121
alternative geophone designs 103–4
Amateur tattoo 270, *270*
American Society for Laser Surgery and Medicine (ASLMS) 200
Anderson, R. R. 268
Andersson-Engels, S. 327
angioplasty, laser *274,* 274–5, **275**
angular vibration velocity, measurement of
  laser torsional or rotational vibrometer *54,* 54–6, *55*
  measurements within rotating system 56, *56*
  periodic speckle noise 57, *57*
  practical considerations 56–8
  speckle noise peak reduction 57, *57*
anti-Stokes Raman scattering 206
applications satellites 187
argon laser trabeculoplasty (ALT) 234
aromatic organic compound *202*
atheroma and angioplasty-like injury 273–4
atmospheric channel 391
  AO technique for downlink and uplink 392
  aperture averaging for downlink 392
  attenuation 391
  coding of downlink and uplink transmitters 392
  mitigating effects of turbulence/scintillation 392
  multi-beaming for uplink 392
  radiance/background light 391
  scintillation/turbulence 391
atom interferometry 176–7
atom–light interactions, theory of
  coherent excitation 129
  incoherent excitation: Einstein A and B coefficients 128
  light shift 129–30
  momentum associated with atom–light interactions: radiation pressure 128–9
  optical bloch equations 130–1
  Rabi solution 130
Auerhammer, J. M. 282
authorized users 357
autocorrelation function (ACF) 29
autodyne frequency tracker 48, *48*
avalanche PDs (APDs) 444
AWG-based optical filters 406–7
Axelrod, D. 329

Bai, K. 39
basic transduction mechanism 103
beam pointing and stabilization 387
beam transport and alignment 284
beat frequency 5
Bell, J. R. 52
Benedict, L. H. 31
Beutler, G. 192
Bigio, I. J. 327
bio-cavity lasers 337–8
biological organization, imaging
  at cellular level 333–4
  of organelles and bacteria 333
  single-molecule detection 333
  at sub-cellular level 333
  in tissue 334
biological samples
  electromagnetic spectrum *322*
  flow cytometry 336–7
  fluorescence correlation spectroscopy 336
  laser probing of (*see* probing of biological samples)
  opto-mechanical manipulation 334–5
  physico-chemical manipulation 335–6
bioluminescence activation 257–8
biomedical laser safety
  classes of lasers in medical use 348–50
  fires 360–1
  hazards to patients 361–2
  incidental hazards 362
  intense pulsed light sources 350
  laser accidents 346–8
  laser safety management 355–8
  laser standards and guidelines 353–5
  precautions 358–9
  principles of quality assurance 350–3
  reflections 359–60
Biomedical Optics (BIOS) Conferences 200
biotechnology
  bio-cavity lasers 337–8
  DNA micro-arrays and genomics 337
  protein micro-arrays and proteomics 337
Birch, D. J. S. 327
blackbody radiation (BBR) 145–6, **146**
blue light hazard (BLH) 350
Bose–Einstein condensate 127
Bragg cells 46, *46,* 88
branch retinal vein occlusion (BRVO) 232
BRET (bioluminescent resonant energy transfer)-PDT 258
broad bandwidth light sources in OCT
  femtosecond lasers 312
  instantaneous linewidth of swept source 316–17
  repetition/sweep rate 317
  source noise 315–16, *316*
  spectrum (central wavelength and bandwidth) 314–15
  supercontinuum sources 313–14
  superluminescent diodes 314
  swept source lasers 313
Bucaro, J. A. 83, 90
Budiansky, B. 89

calcium-stabilized laser 13
calibration 191
cancer treatment, PDT for
  skin tumours 258–9
  tumours of hollow organs 259
  tumours of solid organs 259
carbon dioxide in dermatology 271
carrier frequency 95
cataract surgery 237, *237,* 238*, 238*
Cavendish, H. 178
cavitation 221–2
cavity-stabilization techniques 122–4, *123*
cell phone camera-based fluorescence *303*
centralized (or cloud) radio access network (C-RAN) 432
central serous chorioretinopathy (CSCR) 232, 233
centre of mass corrections 193
charge-coupled device (CCD) camera 31–3, *33,* 62
chemometric techniques 304
chip-to-chip interconnects to trans-oceanic cables *see* fibre-optic transmission systems
choroidal neovascularization (CNV) 232

chromophores or fluorophores 327
Chu, S. 134
classes of lasers in medical use *348,* **348**
  Class 1 349
  Class 2 349
  Class 4 350
  Class 3B 349
  Class 1C 349
  Class 1M 349
  Class 2M 349
  Class 3R 349
clock transition 141, 158
coagulation 267
coated fibres 89–90, *90*
co-channel interference (CCI) 456, *456, 458*
coherent anti-stokes Raman scattering (CARS) 331
coherent excitation 129
cold atom gravimeters *176*
  atomic fountain 176
  atom interferometry 176–7, *177*
  current status of 178
  inertial reference 177
  new methods 178
  Raman transitions and atom interferometery 177
colour modulation techniques 453
communications 365
Compton scattering 128
confocal imaging 301–2
continuous wave (CW) laser 33
cooling, laser
  Doppler cooling 132–4
  light forces 132
  sub-Doppler cooling 134–5
  technology 136–7, *137*
  theory of atom–light interactions 128–31
coordinated multipoint joint transmission (CoMP-JT) 457
Cordiale, P.G. 11
corneal ablation for refractive surgery 236–7
Cramer-Rao Lower Bound (CRLB) 29
Cranch G. A. 1
cross-beam laser velocimeter 48, *48*
cross-correlation cameras *33*
cryogenic optical clocks 152
Cummins, H. Z. 24
cutaneous vascular lesions 268–9
CW laser Doppler blood flow measurements 307
Czarske, J. 29

Dahlstrom C. E. 10
Dailey, M. E. 328
datacentre interconnects 433–4
data screening 189
3D-deformations, single loading *74,* 74–5, *75*
deep neural network (DNN) 464
Defence Evaluation and Research Agency (DERA) 94
  hydrophone designs *94*
denaturation 267
density operator 130
dermatology
  carbon dioxide 271
  cutaneous vascular lesions 268–269
  Erbium:YAG laser 271–2
  hair, laser treatment of 270–1
  photothermal interactions *267,* 267–8, *268*
  pigmented lesions and tattoos, laser treatment of 270
  selective photothermolysis 268–9
  treatment of PWSs 268–9
detectors 185
Deutsch, T. F. 283
DGV or planar Doppler velocimetry (PDV) 24
diabetic macular edema (DME) 232
dielectric breakdown 237
diffuse lamellar keratitis (DLK) 245
diffuse optical spectroscopy *293,* 293–8, *294, 295, 296, 297*
diffuse optical tomography (DOT) 209
digital image correlation (DIC) 61
digital modulation techniques 450–1
  colour modulation techniques 453
  and multiplexing 428–9
  OFDM-based MCM 451–3
  in optical communications, practical aspects 429–30
  single-carrier modulation techniques 451
digital speckle pattern interferometry (DSPI) 61
digital techniques 98
directional couplers 88
directly modulated distributed feedback (DFB) 431–2
'discrete ordinates,' method of 210
discrete power level stepping 455
DNA micro-arrays and genomics 337
Doppler cooling 133
  equilibrium temperature 133
  one-dimensional 132–3
  optical molasses in three dimensions 134
  two counter-propagating beams 132–3
Doppler global velocimetry (DGV) 24, 34–5, *35*
Doppler model *25,* 25–6, *26*
Doppler signal processing 48–9
3D Optical Lattice Clocks 151–2
double-balanced mixer (DBM) 12
double-plane mirror interferometers 7, *7*
double-pulse method 77
Dragone, C. 406
dressed-state Hamiltonian 129
dual-beam arrangement 24
dynamic measurements, ESPI
  applications of 77
  double-pulse method 77
  stroboscopic method 76, *77*
  time-averaged method 75–6, *76*

echelle diffraction gratings (EDG) 406
Edwards, C. S. 282, 283
Einstein A and B coefficients 128
Einstein's photon recoil concept 128
Ekins, R. P. 337
elastic scattering spectroscopy 203, *204,* 298–9
electro-absorption modulators (EAMs) 432
electronic speckle pattern interferometry (ESPI)
  description of technique 62–3
  dynamic measurements 75–7
  evaluation of the interference phase 65–70
  Fringe interpretation 63–4
  out-of-plane and in-plane deformation measurement *64,* 64–5
  principle of *62*
  spatial phase-shift 73–5
  temporal phase-shift 70–3
electro-optic phase modulators (EOMs) 6
electro-optic switches 88
endogenous molecular probes 324–5
  endogenous absorbers 324
  endogenous fluorophores 324–5
endoscopes, flexible
  cancer of lungs 263
  cancers of gastrointestinal tract 262
  urology 263
endothelial keratoplasty 237
Endpoint Management (EpM) 233, *233*
entangled clocks 152
epidermal melanin 269
equipment 358
erbium-doped fibre amplifiers (EDFA) 85
Erbium:YAG laser in dermatology 271–2, *272*
Ewins, D. J. 52
excimer laser angioplasty
  coronary arteries 275
  peripheral (femoral) arteries 275–6
excimer laser-based procedures 243–4
exogenous probes 325–6
  fluorescent protein-based probes 325–6
  quantum dots 326
  Raman probes—surface-enhanced Raman scattering 326
  small organic fluorescent probes 325
exotic atomic clocks 152
extended cavity diode laser (ECDL) configuration 13
eye, optical properties of 228, *229*

Fabry–Perot cavity determination 179–80
Fabry–Perot interferometers 15
Fabry–Perot resonator (FPI) 13
falling corner cube gravimeters 174, *175*
  current status and future work 175–6
  drag-free chamber and drop mechanism 175
  falling corner cube 175
  laser source 175
  signal acquisition and analysis 175
  super-spring inertial reference. 175
far-off resonance optical traps 136
Fast Fourier Transform (FFT) 311
FBG laser sensor *102*
"feeder vessels" 257
femtosecond combs 165–6
  absolute optical frequency measurement 166, *166*
  optical frequency ratio measurement 166–7, *167*
  optical frequency synthesis 167
femtosecond laser (FSL) technology 243, 312
  Assisted Stromal Lenticule Addition Keratoplasty (SLAK) 246
  based procedures 245
femtosecond lenticule extraction (Flex) 243
Fermi gas 127
Fibre Bragg gratings (FBG) 88
fibre laser sensors 102

fibre-optic communication systems
  datacentre interconnects 433–4
  directly modulated distributed feedback (DFB) 431–2
  electro-absorption modulators (EAMs) 432
  metro and long-haul networks 434–9, *435–8*
  mobile front-haul and PONs 432–3
  on-off keying (OOK) *430*
  role of 427–8, *428*
  short-reach systems 430–2, *431, 432*
  sub-marine transmission *439,* 439–40
fibre-optic transmission systems
  communication systems from short-reach to ultra-long-haul (*see* fibre-optic communication systems)
  fused-silica optical fibre 427
  modulation and multiplexing 428–30
  radio-frequency (RF) communications 427
  role of fibre optics in communications 427–8
fibre-to-the-chip *see* vertical cavity surface-emitting laser (VCSELs), VLSI circuits
fires 360–1
FLARM 191
FLEx procedure 246
flight laser transmitter 391
flow cytometry 336–7
fluorescence correlation spectroscopy 306, 336
fluorescence quantum yield 203
fluorescence spectroscopy and imaging 302–3
fluorescent protein-based probes 325–6
fluorophores 323
Fourier domain OCT (FD-OCT) 310–12
Fourier transformation 68–70
fractional frequency reuse (FFR) 457
free beam delivery 284
free-electron laser (FEL) *282*
  beam transport and alignment 284
  clinical application of FEL 283–4
  laser ablation 282–3
  neurosurgery 285
  ophthalmology 285–6
  pump-probe imaging *283*
  surgical beam delivery 284–5
frequency chains 4
frequency domain multiplexing (FDM) *98,* 98–9
frequency stabilization
  advantages of 112
  based Doppler-free spectroscopy 116–18
  based on Doppler-limited absorption 116
  of He-Ne laser to gain curve *115,* 115–16
  and reproducibility, evaluation of *121,* 121–2, *122,* **122**
frequency stabilized lasers
  determination of fundamental constants 114–15
  for dimensional interferometry 112–14
  frequency-stabilized diode lasers *13,* 13–14
  He-Ne laser stabilized to gain profile *9,* 9–10
  interferometry for gravitational wave detection 114
  iodine-stabilized He-Ne laser at λ = 633 nm *10,* 10–11, *11*
  referenced to iodine, rubidium and acetylene 118–21, *119, 120*
  stabilized frequency-doubled Nd:YAG laser at 532 nm wavelength 11–12, *12*
  wavelength or frequency standards **8,** 8–9
Fried's parameter 391
Fringe formation, ESPI
  phase change and deformation components, relationship between 63–4
  principle of 63, *63*
Fringe model *24,* 25
fundamental interactions
  absorption 202–3
  elastic scattering 203, *204*
  inelastic scattering 205–6, *206*
  luminescence 203
  miscellaneous 206–7
  scattering by dielectric particle *204,* 204–5, *205*
  scattering by many particles 205
fundamental length metrology
  evolution of metre, definition and realization 3–4
  frequency stabilized lasers 8–9
  grating interferometers 7–8
  interpolation 14–15
  laser interferometers (*see* laser interferometry)
  multiple wavelength interferometry 18–19
  refractometry 14
  SI-unit of length, definition of *4*
fused-silica optical fibre 427

gas-cell-absorption-based stabilization techniques
  Doppler-free spectroscopy 116–18
  Doppler-limited absorption 116
  He-Ne laser to gain curve 115–16
  referenced to iodine, rubidium and acetylene 118–21
gastrointestinal tract, cancers of 262
generalized spatial modulation (GSM) 454–5
geodetic spheres 186–7
Gerstenkorn, S. 8
Gill, E. 192
glare points 36
global navigation satellite systems (GNSS) 187, 195
global network 191–2
Global Positioning System (GPS) 195
Golf ball velocity of deformation *53*
Gourley, M. F. 337
gradient force or dipole force 131
grating-based optical filters 410, *411, 412*
grating interferometers 7, *8*
gravitational wave detection
  Fabry-Perot cavities 173–4
  first detection 172
  future work on Earth and in space 174
  gravitational wave signal extraction 174
  interferometer control 174
  laser source and input optics 172–3
gravitation measurements with laser interferometry; *see also* laser interferometry
  absolute gravimetry 174–8
  determination of Newtonian constant of gravitation 178–9
  gravitational wave detection 172–4
Gray mapping 430
green fluorescent protein (GFP) 325
grid photocoagulation 232
ground segment
  detectors 185
  lasers 183
  telescopes 183–5
  timers 185–6
ground station
  ground PAT 393
  ground receive telescope 392
  ground uplink transmitter 392
Gruneisen coefficient 219
Gruneisen parameter 306

hair, laser treatment of **270,** 270–1
Halliwell, N. A. 1
Hanes, G. R. 10
hazards
  to eye 358
  incidental 362
  to patients 361–2
  to skin 358–9
heat shock protein (HSP) 233
Hebbian learning 470
heterodyne interrogation scheme 95–6
heterodyne laser interferometry 5–7, *6,* 84
Heydemann, P. L. M. 15
Hockham, G. A. 365
hollow organs, tumours of 259
hollow waveguides (HWG) 284–5
holographic gratings (HGs) 13
homodyne laser interferometry 4–5, *5,* 84
homodyne phase-generated carrier (PGC) 94–5, *95*
HoYAG laser ENT application *347*
hyper-Ramsey interrogation 161
hyperspectral 327

illumination angle 62
imaging, laser-based
  levels of biological organization 333–4
  properties of light (*see* properties of light, imaging)
incidental hazards
  carcinogenic dyes 362
  electrical danger 362
  fire 362
  laser plume 362
incident laser beams 55
incident reports 358
incoherent excitation 128
index guide 370
indoor VLC channel characterizations
  comparison of NLoS channel calculation methods 450
  deterministic method 448
  efficient analytical method 449–50
  frequency domain calculation 448
  line-of-sight channel and non-line-of-sight channel 447–8
  Monte Carlo method 448
  simplified sphere model 448–9
inelastic scattering 205–6

in-fibre laser (IFL) 102
infrared transmitting fibre-optics 285, *285*
in-line Michelson 99
input-light polarization control techniques 102
intense pulsed light sources (IPLs) 350, *350*
interference phase, ESPI
  spatial phase-shift technique 67–70
  temporal phase-shift technique 65–7
interferometric hydrophone basic principles 83–4
International Association of Geodesy (IAG) 191
International Laser Ranging Service (ILRS) 182, **187,** *192*
internationally recommended optical frequency standards 167–8, **168**
interpolation 14–15
interstitial laser photocoagulation (ILP) 263–5, **264,** *264*
inter-system crossing 203
'inverse problem' 297
in vivo methods and applications
  confocal imaging 301–2
  CW laser Doppler blood flow measurements 307
  diffuse optical spectroscopy and spectral imaging 293–8
  elastic scattering spectroscopy 298–9
  fluorescence correlation spectroscopy 306
  fluorescence spectroscopy and imaging 302–3
  non-linear techniques 307
  optical coherence tomography 298–301
  optical imaging in pathology 307
  photoacoustic imaging 305–6
  photothermal spectroscopy 306–7
  polarimetry 307
  Raman spectroscopy and imaging 303–5
  speckle interferometry 306
  terahertz imaging 307
  white-light (reflectance) imaging 292–3
ionizing radiation activation 258
ion optical clocks
  operation 147–8, *148*
  referenced to quadrupole transitions *160,* 160–1, *161*
  referenced to transitions other than quadrupole transitions *161,* 161–2
  systematics 148, **148**
ion-trapping technique 159–60, *160*
Ishijima, A. 329, 333
Izatt, J.A. 315

Jansen, E. D. 199
Jenkins, M. P. 277
Juhasz, T. 237

Kao, K. C. 365
Katzir, A. 285
Katz, J. 39
Kaula, W. M. 192
Keck, W. M. 281, 283
keratomileusis 236
Kitamura, K. 329
Knudsen, S. 91, 93
Koster, R. 275
Krueger, R. R. 238

Lagakos, N. 90
LAGEOS 181, 185, 188
Lakowicz, J. R. 325, 327
Laming, R. I. 48
LARES 187, *187,* 195
large-scale array architectures 101
laser ablation
  ablation with FEL 282–3
  pulse structure effects 283
  wavelength effects 282–3
laser-activated processes 336
  chromophore-assisted laser inactivation (CALI) 336
  fluorescence recovery after photobleaching (FRAP) 336
  uncaging 336
laser-assisted in situ keratomileusis 237
laser beam pointing 386–7
laser capture micro-dissection 307
lasercom
  benefits 384
  challenges 384–5
laser controlled area 357, *357*
laser diodes (LDs) 7, 444
laser Doppler anemometry (LDA) 24
laser Doppler velocimetry (LDV)
  data processing 30–1
  Doppler model *25,* 25–6, *26*
  dual-beam optical arrangement 24, *24*
  Fringe model *24,* 25
  local light intensity *24*
  receiving optics 27, *27*
  signal processing *29,* 29–30, *30*
  system configurations *27,* 27–9, *28*
  transmitting optics 26, *26*
laser heating 215
Laser Interferometer Gravitational Wave Antenna (LIGO) interferometer 18, 172, *172, 173*
Laser Interferometer Space ntenna (LISA) 17
laser interferometry 4
  accuracy limits of 15–16, *16*
  applications of 16–18, *17*
  heterodyne 5–7, *6*
  homodyne 4–5, *5*
  set-ups 7
Laser Link (T2L2) experiment 195–6
laser protection adviser 355
laser protection supervisor 355–6
laser scanning confocal fluorescence microscopy (LSCFM) 328, *328, 330*
laser speckle effects *49,* 49–51, *50, 51*
laser stabilization for precision measurements
  basic spatial and spectral characteristics of lasers 111–12
  cavity-stabilization techniques 122–4
  evaluation of frequency stability and reproducibility 121–2
  frequency stabilization, advantages of 112
  frequency-stabilized lasers, applications of 112–15
  gas-cell-absorption-based stabilization techniques 115–21
laser torsional or rotational vibrometer 54–6
laser trabeculoplasty 234
laser velocimetry
  application issues 38–9, *39,* **40**
  Doppler global velocimetry 34–5
  for fluid flow measurements *23*
  future directions 39–40
  laser Doppler velocimetry 24–31
  particle image velocimetry 31–4
  PD techniques 35–8
laser vibrometers 44
  laser speckle effects 49–51
  measurement of angular vibration velocity 54–8
  measurements on rotating targets *51,* 51–2, *52*
  principles of *44,* 44–5, *45*
  scanning, impact measurements and practical considerations 52–3
  solid-surface vibration measurements *45,* 45–9
  vibration velocity contours *43*
LASIK procedure 243, *244,* 245, 246
LASSO (Laser Synchronization from Stationary Orbit) experiment 195
Latina, M. A. 235
Launch optics arrangement *100*
legislative health and safety requirements 358
Lehmann, B. 24
Lewin, A. C. 47
LiFi (Light-fidelity) 444; *see also* visible light communications (VLC)
  access point layout and SINR characterizations 456–7
  adaptation of MIMO techniques (*see* multiple-input multiple-output (MIMO) techniques)
  attocell network 456
  electrical-to-optical conversion based on multiple light sources 455
  interference mitigation techniques 457–8
  MIMO optical source and detector deployment 455
  repetition coding 454
  spatial modulation 454–5
  spatial multiplexing 454
light and matter interaction
  absorption of light 323, *323*
  emission of light 323
  quantum confinement—quantum dots 324
  scattering of light 323–4
light delivery systems 252–5, *253, 254*
light emitting diodes (LEDs) 444
light forces
  deceleration of an atomic beam 132
  dipole force 132
  spontaneous force 132
light sources
  lasers 250–2, *251, 252*
  non-laser sources 252, *252*
light–tissue interactions
  fundamental interactions 202–7
  optical properties 207–9
  physical models of light propagation 210–12
  therapeutic laser–tissue interactions 213–24
linear inviscid force equation 93
linear optical responses

  - laser microscopy beyond the diffraction limit 328–9
  - laser scanning confocal fluorescence microscopy (LSCFM) 328
- link budget **385**
  - optical channel 385
  - receive channel 385–6
  - transmit channel 385
- Littman, M. 13
- local field corrections 32
- low Earth orbiters (LEOs) 184
- Luc, P. 8
- luminescence 203, *203*
- lunar laser ranging (LLR) 196
- lungs, cancer of 263, *263*

- Mach–Zehnder modulators (MZMs) 433
- Mach–Zender interferometer 47, *47, 83,* 84–5
- macula, non-damaging laser therapy of 232–3
- magic-trapping frequency 162
- magnetic and optical dipole traps 136
- magneto-optical trap (MOT) *135,* 135–6
- mandrel hydrophones 90–2, *91*
- master-oscillator-power-amplifier (MOPA) 391, *391*
- matter wave interferometers 172
- Maxwell, J. C. 445
- measurement chain 16
- medical diagnostics
  - comparison of techniques and general future trends 307–8
  - *in vivo* methods (*see in vivo* methods and applications)
  - optical wavelengths 291–2
  - UV-vis-NIR photon energies 291
- medical examination 357–8
- medical, lasers in 199–200
- Melling, A. 39
- memcomputing 464–5
- memelement 465
- metro and long-haul networks 434–9
- metrology, laser in 1
- Meyer-Schwickerath, G. 228
- Michelson interferometer *44,* 45, *45, 46, 84,* 85, *113*
- micro-channel plate (MCP) detectors 185
- micropulse laser 233
- micro-sized LEDs 445–6
- ‘micro-wave orbits’ 195
- Mie Theory model 298–9
- Miles, T. J. 57
- Misura Accurata di Gmediante Interferometria Atomica (MAGIA) experiment 179
- mobile front-haul and PONs 432–3
- modified articulated arm 284
- modulation 390
  - and coding 390–1
- modulation index 117
- Molkenstruck, W. 46
- Monte Carlo Simulation *210,* 210–11, *212,* 448
- Montenbruck, O. 192
- Moore’s law *464*
- Mourant, J. R. 327
- MRR-based optical filters *407*
  - with box-like responses 408–9
  - with large FSRs 408
  - polarization-selective MRR optical filters *409,* 409–10, *410*
  - ultra-compact MRRs on Silicon **408**
- Mueller matrix 205
- multi-carrier modulation (MCM) technique 450
- ‘multi-flux method’ 210
- multimode interference (MMI) 402, *403*
- multi-mode silicon photonics 402–3
  - on-chip mode (de)multiplexers 403–4
  - sharp multi-mode waveguide bends *405,* 405–6
- multi-photon laser scanning fluorescence microscopy 330
- multi-pixel calculation 67–8
- multiple-input multiple-output (MIMO) techniques 454–6
  - adaptation of 456
  - electrical-to-optical conversion based on multiple light sources 455
  - optical source and detector deployment 455
  - repetition coding 454
  - spatial modulation 454–5
  - spatial multiplexing 454
- multiple quantum well (MQW) 368
- multiple scattered light 205
- multiple wavelength interferometry
  - gauge block calibration 18–19
  - interferometric distance measurements 19
- multiplexed FBG laser sensors *103*

- nanocrystals *see* quantum dots
- nanosurgery (laser scalpels) 335
- Nash P. J. 1
- navigation constellations 187–8
- Nd:YAG (neodymium:yttrium-aluminium-garnet) laser, 33, 237, *237*
- Nd:YLF lasers 33
- ‘negative velocity potential’ 220
- neointimal hyperplasia (NIH) 273
- neural networks 304
- neurosurgery 285, *286*
- neutral atoms, optical clocks 162–3
  - optical lattice clock operation 148–50
  - optical lattice clock systematics 150–1
  - preparation of neutral atoms for interrogation 163–4
  - spectroscopy and laser locking 164, *164*
  - systematic evaluations 164–5
- Newtonian constant of gravitation
  - atom interferometer determinations 179
  - free-falling corner cube determination 179
  - suspended Fabry-Perot cavity determination 179–80
- next-generation clocks
  - cryogenic optical clocks 152
  - 3 D Optical Lattice Clocks 151–2
  - entangled clocks 152
  - exotic atomic clocks 152
  - superradiant optical clocks 152
- NLoS channel calculation methods, comparison of *449,* 450, **450,** *450*
- noise properties
  - laser frequency noise 86–7, *87*
  - laser intensity noise 86
- noise sources and phase resolution 96–7, *97*
- nominal ocular hazard distance (NOHD) 357
- non-damaging laser therapy of the macula 232–3
- non-damaging retinal laser therapy (NRT) 234
- non-destructive-test 70–1
- non-linear interference noise (NLIN) 438
- non-linear optical responses
  - coherent anti-stokes Raman scattering 331
  - multi-photon laser scanning fluorescence microscopy 330
  - second harmonic generation 330–1
  - surface-enhanced Raman scattering 331
- non-linear techniques 307
- non-volatile photonic memory 465–7
- non-von Neumann architecture
  - memcomputing 464–5
  - photonic memelements for 467–8
- normal point (NP) data 193

- ocular phototherapy 228
- OFDM-based MCM 451–3
  - DCO-OFDM 451–2, *452*
  - hybrid OFDM techniques 453
  - unipolar OFDM techniques 452–3, *453*
- Oheim, M. 329
- on-chip mode (de)multiplexers 403–4
- on-chip polarization-handling
  - silicon PBSs 400–1, *401*
  - silicon PRs *401,* 401–2
- one A-scan 309
- ophthalmic laser therapy and surgery
  - early history 228
  - future directions 239
  - optical properties of eye 228
  - photocoagulation 229–30
  - photomechanical interactions 235–7
  - photothermal therapy 228, *229*
  - tissue-selective therapy 234–5
  - transparent tissue surgery with ultrashort-pulse lasers 237–9
- ophthalmology 285–6
- optical add/drop multiplexers (OADM) 101
- optical associative memory (OAM) 468–70, *470*
- optical atomic clocks *158*
  - absolute optical frequency measurement 166
  - atomic parameters **159, 163**
  - based on neutral atoms 162–3
  - based on single trapped ion
  - blackbody radiation 145–6
  - clock accuracy 141–2
  - clock interrogation sequences 144
  - clock stability *140,* 140–1
  - Doppler shifts 146
  - electric fields 145
  - gravitational redshift 146–7
  - internationally recommended optical frequency standards 167–8
  - magnetic fields 145
  - many-body quantum physics 153
  - neutral atom optical lattice clocks 148–51
  - next-generation clocks 151–2
  - principles of 139–40
  - single-ion optical clocks 147–8
- optical beam specification 355
- optical Bloch equations 130–1

- optical coherence tomography (OCT) 298–301, *310*
  - broadband sources (*see* broad bandwidth light sources in OCT)
  - confocal imaging *301*
  - fibre-optic *299*
  - Fourier domain 310–12
  - principle of *298*
  - time domain 310
- optical data storage
  - associative optical memory for pattern recognition 468–70
  - method of optical memory 468
  - non-volatile photonic memory 465–7
  - non-von Neumann architecture 464–5, 467–8
  - von Neumann architecture 464
- optical fibre hydrophones
  - application studies 104–6, *105, 106*
  - basic transduction mechanism 88–9
  - brief history 83
  - coated fibres 89–90
  - at higher frequencies and directionality 93
  - interferometry (*see* optical fibre interferometry)
  - mandrel hydrophones 90–2
  - optical geophone 103–4
  - optical systems and multiplexing 98–102
  - practical hydrophone designs 93–4, **94,** *94*
  - responsivity with depth *92,* 92–3
  - sensitivity to other effects 93
  - signal processing 96–8
  - sonar system requirements **82,** 82–3
- optical fibre interferometry
  - acousto-optic modulators or Bragg cells 88
  - amplitude modulators 88
  - characteristics of lasers used in sensors 87, **87**
  - configurations 84–5
  - directional couplers 88
  - Fibre Bragg gratings (FBG) 88
  - interferometric hydrophone basic principles 83–4
  - lasers 85–6
  - Mach–Zender interferometer *83,* 84–5
  - Michelson interferometer *84,* 85
  - modulation properties 86
  - noise properties 86–7
  - optical receivers 87
  - phase modulators 88
  - Sagnac interferometer *84,* 85
  - semiconductor optical amplifiers 88
  - wavelength division multiplexing (WDM) devices 88
- optical frequency combs 142, *143,* 144
- optical frequency synthesis 167
  - comb 165
  - divider 167
  - ratio measurement 166–7
- optical geometry/interferometer 47–8
- optical geophone 104, *104*
  - alternative geophone designs 103–4
  - basic transduction mechanism 103
  - system configurations for geophone use 404
- optical heterodyning 25
- optical lattice clock
  - operation 148–50, *149, 150*
  - systematics *150,* 150–1, **151**
- optical monitoring and dosimetry 255–6, *256*
- optical parametric oscillators (OPO) 199, 281
- optical penetration depth 215
- optical radiation safety
  - documents 353–5
  - policy 355
- optical safety guidance 355
- optical spectroscopy
  - applications to tissue diagnostics 327
  - steady-state measurements 326
  - time-resolved measurements 326–7
- optical switches 410–11
- optical systems and multiplexing
  - FDM technique 98–9
  - fibre laser sensors 102
  - input-light polarization control techniques 102
  - large-scale array architectures 101, *101*
  - optical hydrophone technology 102
  - overcoming polarization-induced signal fading 101–2
  - received-light SOP selection techniques 102
  - TDM architectures analysis 99–100
- optical tweezers 128, 213, 334–5
- optical wavelengths 291–2
- ‘optical window’ 292
- ‘Optoelectronic VLSI’ (OE-VLSI) technology 368
- opto-mechanical manipulation of biological samples
  - nanosurgery (laser scalpels) 335
  - optical tweezers 334–5
  - perforation (optoporation) 335
  - tissue welding and soldering 335
- optoporation *see* perforation
- organic LEDs 446
- orthogonal laser vibrometers 51–2, *52*
- out-of-plane and in-plane deformation measurement 64–5
  - under continuous loading 73–4, *74*
  - temporal phase-shift technique, ESPI 70–1
- out-of-plane spatial phase-shift ESPI system 69

- panretinal photocoagulation (PRP) 230
- Parrish, J. A. 268
- particle image velocimetry (PIV) 24, *32*
  - basic principles 31–2
  - choice of laser 33
  - three-velocity component PIV *33,* 33–4, *34*
- particle tracking velocimetry (PTV) 24, *32*
- passive interrogation schemes 94–6
  - heterodyne interrogation scheme 95–6
  - homodyne phase-generated carrier (PGC) 94–5, *95*
- passive silicon photonic integrated components and circuits
  - future trends 414–15
  - multi-mode silicon photonics 402–3
  - reconfigurable silicon photonic devices and circuits 410–14
  - silicon nanophotonic waveguides *398,* 398–9, *399*
  - silicon photonic devices for on-chip polarization-handling 400–2
  - wavelength-selective silicon photonic devices 406–10
- pathology, optical imaging in 307
- patients, hazards to
  - carbonization 362
  - endotracheal tube ignition *361,* 361–2
  - misdirected beam 361
  - risk to eyes 361
  - risk to skin 361
- patterned laser trabeculoplasty (PLT) 234
- pattern recognition 468–70
- pattern-scanning retinal photocoagulation 232
- perforation 335
- Pfeifer, H. J. 24
- phase change materials (PCM) 465–8, *466, 467, 468*
- phase Doppler (PD) techniques 24
  - basics of light scattering 35, *36*
  - implementation 38, *38*
  - measurement principle 36–7, *37*
- phase modulation technique 13
- phase modulators 88
- phase-sensitive detection (PSD) 10, 117
- phosphorescence 203
- photoablation 236
- photoacoustic imaging (PAI) 305–6
- photocoagulation 229–30, *230*
  - laser trabeculoplasty 234
  - non-damaging laser therapy of macula 232–3
  - optimization of pulse duration 230–2, *232*
  - pattern-scanning retinal photocoagulation 232
  - real-time monitoring of tissue temperature 234
  - retinal plasticity 230, *231*
- photodetector output spectrum 45, *45*
- photodynamic therapy (PDT) **264**
  - alternative photosensitizer activation schemes 257–8
  - applications of 259
  - in arteries 276–8
  - for cancer treatment 258–9
  - clinical data 277–8, *278*
  - of injured and diseased arteries 276, *277*
  - light-activated chemotherapy for treatment of cancer *214*
  - light delivery systems 252–5, 276–7
  - light sources 250–2
  - optical monitoring and dosimetry 255–6
- photomechanical effects
  - plasma formation 224
  - spallation and cavitation 221–2
  - thermoelastic expansion 219–21, *221, 222*
  - vaporization 222–4
- photomechanical interactions 235
  - corneal ablation for refractive surgery 236–7
  - selective laser trabeculoplasty 235–6
  - selective RPE therapy 235
- photorefractive keratectomy (PRK) 236, 243
- photosensitizer activation schemes
  - bioluminescence activation 257–8
  - ionizing radiation activation 258

  two-photon activation 257
  ultrasound activation 258
  up-converting nanoparticles 257
photosensitizers 234
photothermal effects
  laser heating 215
  thermal coagulation parameters **218**
  thermal damage 216–19, *217, 218*
  thermal diffusion 215–16, *216*
photothermal interactions 267–8
photothermal spectroscopy 306–7
photothermal therapy
  photothermal interactions 228, *229*
  quantification of thermal damage 229
physico-chemical manipulation of biological samples
  laser-activated processes 336
  pump-probe technique 335–6
pigmented lesions and tattoos, laser treatment of 270
PIN-PDs and APDs 446–7
Planck's law 145
plasma formation 224
polarimetry 307
polarization optics 46
polarizing beamsplitter (PBS1) 47
polygon-scanner-based filter *313*
popcorn effect 223
port-wine stain (PWS), treatment of 268–9, *269*
positive-intrinsic-negative photodiodes (PIN-PDs) 444
'Pound–Drever–Hall' method 123
power spectral density (PSD) 29
practical hydrophone designs 93–4
precautions
  equipment features 359
  hazards to eye 358
  hazards to skin 358–9
  protective eyewear 359
  risk assessment 359
  room layout 359
pre-coding and post-processing 454
Principal Component Analysis 304
probing of biological samples
  endogenous molecular probes 324–5
  exogenous probes 325–6
  optical spectroscopy 326–7
  spectral imaging 327–8
properties of light, imaging
  linear optical responses 328–9
  non-linear optical responses 330–1
  spectral-based applications 331–3
protein micro-arrays and proteomics 337
pseudotumour cerebri 285
'pseudo vibrations' 50
pulse amplitude modulation (PAM) 450
pulsed ESPI technique 77
pulse position modulation (PPM) 450
pulse structure effects 283
pulse width modulation (PWM) 450
pump-probe technique 335–6
PV solar cells 447

Q-switched Nd:YAG laser 270, *270*
quality assurance 351
  CO_2 laser 352, *352*
  principles of 350–3, *351*
quantum dots 324, 326

Rabi solution 130
radiation pressure 128
radiation pressure force or spontaneous force 131
radio-frequency (RF)
  communications 427
  Paul trap oscillating electric fields 147, *147*
radiometer experiments **128**
Ragni, D. 40
Raman probes 326
Raman spectroscopy *206, 303*
  and imaging 303–5
Raman transitions 177
ranging policy 189–91
Rayleigh scattering 323
received-light SOP selection techniques 102
receivers
  PIN-PDs and APDs 446–7
  PV solar cells 447
  single-photon APDs 447
receiving optics 27, *27*
recoil limit 135
reconfigurable silicon photonic devices and circuits
  optical switches 410–11
  ROADMs 413–14
reference beam mode 24
reflections 359–60, *360*
refractive index (RI) 186
refractive surgery 237
  complications 244–5
  excimer laser-based procedures 243–4
  FSL-based procedures 245
  newer application of small-incision lenticule extraction 246
  ReLEx procedures 246
refractometry 14, *15*
Reibold, R. 46
ReLEx procedures 246
repetition coding 454
repetition/sweep rate 317
resonant cavity LEDs 446
resource partitioning 457
restenosis 273
retinal plasticity 230
retroreflectors 186
Richards-Kortum, R. 327
risk assessment 359
ROADMs *413,* 413–14, *414*
Rockson, S. G. 277
Rothberg, S. J. 52

safety management, laser
  authorized users 357
  equipment 358
  equipment modification 358
  incident reports 358
  laser controlled area 357
  laser protection adviser 355
  laser protection supervisor 355–6
  legislative health and safety requirements 358
  local rules 356–7
  medical examination 357–8
  optical radiation safety policy 355
  training 357
Sagnac interferometer *84,* 85, *85*
satellite laser ranging (SLR)
  accuracy and quality control 193
  calibration 191
  centre of mass corrections 193
  data screening 189
  future trends 196
  global network 191–2
  ground segment 183–6
  lunar laser ranging 196
  ranging policy 189–91
  safety 191
  signal strength 188, *189*
  space debris 195
  space segment 186–8
  support of scientific missions 194–5
  terrestrial reference frame 193–4
  time transfer 195–6
  track detection 188–9, *190*
  tracking of GNSS constellations 195
  working principle, description 182, *182*
satellite missions 186
saturation 128
scattering by dielectric particle 204–5
scattering of light 323–4
Scheimpflug configuration 33
Schmoranzer, J. 329
Schroder, A. 40
secondary cataract 237
second harmonic generation (SHG) 330–1
Selbach, H. 47
selective laser trabeculoplasty (SLT) 235
selective photothermolysis 268; *see also* dermatology
  beam diameter 269
  epidermal melanin 269
  laser treatment of hair by 270–1
  shadowing effects 269
selective RPE therapy (SRT) 235, *236*
semiconductor optical amplifiers 88
semiconductor optical amplifiers (SOAs) 313
sensor coordinate system 71
shadowing effects 269
"Shake-the-Box" 40
Shapiro, H. M. 337
short-reach systems 430–2
signal detection 389–90
signal processing
  digital techniques 98
  dynamic range 97–8
  noise from optical amplifiers 97
  noise sources and phase resolution 96–7
  passive interrogation schemes 94–6
signal-to-noise ratio (SNR) 29
silicon nanophotonic waveguides *398,* 398–9, *399*
simplified sphere model 448–9
single-carrier modulation (SCM) techniques 450, 451

single-ion optical clocks
  ion clock operation 147–8
  ion clock systematics 148
single-photon avalanche diodes (SPADs) 185, 447
single-photon ranging 188
Singlet Oxygen Luminescence Dosimetry (SOLD) 255
single trapped ion, optical clocks
  evaluation of systematic shifts 162, **162**
  ion optical clocks 160–2
  ion trapping 159–60
Sisyphus cooling 135
SI Unit Definitions, the World Clock, Navigation, and Geodesy 152–3
skin tumours 258–9
slot correlation 31
small-incision lenticule extraction 246
small organic fluorescent probes 325
SMILE procedures 243, 246
Smit, M. K. 406
solid organs, tumours of 259
solid-surface vibration measurements
  Doppler signal processing 48–9
  frequency-shifting and/or direction ambiguity removal *45,* 45–6, *46*
  optical geometry/interferometer 47–8
  solid-surface scattering of laser light: laser speckle 49
sonar system requirements 82–3
Sonodynamic Therapy 258
space debris 195
space segment
  applications satellites 187
  geodetic spheres 186–7
  navigation constellations 187–8
  retroreflectors 186
  satellite missions 186
spallation 221–2
spatial acquisition 387
spatial light modulation imaging 297
spatial modulation 454–5
spatial multiplexing 454
spatial phase-shift technique, ESPI
  carrier-frequency and Fourier transformation *68,* 68–70
  3D-deformations under single loading 74–5
  multi-pixel calculation 67–8, *68*
  out-of-plane and in-plane deformation measurement under continuous loading 73–4
spatial summing 455
speckle interferometry 306
spectral-based applications
  with lasers operating in 'water-window' region 332–3
  with soft X-rays in 'water-window' region 331–2, *332*
spectral domain OCT (SD-OCT) 311, *311*
spectral imaging 293–8, 327–8
Srinivasan, R. 236
stability region 160
Stanbridge, A. B. 52
standard quantum limit (SQL) 18
standards and guidelines, laser
  electrical and mechanical construction 353
  general optical safety guidance 355
  International Electrotechnical Commission (IEC) *354*
  International Standards Organizations **353**
  optical beam specification 355
  optical radiation safety documents 353–5
STARLETTE 181, 187
states of polarization (SOPs) 101
steady-state measurements 326
Stella 187
stimulated emission depletion microscopy 329
Stokes beam 305
Stokes Raman scattering 206
stress confinement 219
stroboscopic method 76, *77*
sub-Doppler cooling *134,* 134–5
sub-marine transmission 439–40
supercontinuum sources 313–14
superluminescent diodes (SLD) 314, *314, 316*
superradiant optical clocks 152
support of scientific missions 194–5
surface-enhanced Raman scattering (SERS) *305,* 326, 331
surgical beam delivery
  free beam delivery 284
  hollow waveguides (HWG) 284–5
  infrared transmitting fibre-optics 285
  modified articulated arm 284
survival weighting 211
swept-source OCT (SS-OCT) 311, *312,* 313, *313*
  instantaneous linewidth of 316–17
switching voltage 433
synthetic wavelength 19
system configurations for geophone use 404

Takahashi, H. 406
technology maturity 385
telescopes 183–5
temporal phase-shift technique, ESPI
  4 +1 Algorithm 66
  4 +4 Algorithm 65–6
  generation of phase-shift 67, *67*
  measurement of three-dimensional deformations and strain 71–3, *72, 73*
  other algorithms 66–7
  out-of-plane and in-plane deformation measurement and non-destructive-test *70,* 70–1
terahertz imaging 307
terrestrial reference frame 193–4
therapeutic laser–tissue interactions *213*
  photochemical effects 213–14
  photomechanical effects 221–4
  photothermal effects 214–15
thermal blooming 230
thermal confinement 235
thermal damage 216–19
  quantification of 229
thermal diffusion 215–16
thermoelastic expansion 219–21
third harmonic technique or modulation technique 11
time-averaged method 75–6
time division multiplexed (TDM) architectures 98
  analysis 99–100, **100**
  progressive ladder architecture (PLA) 99–100
  pulsed reflectometric architecture 99
time domain OCT (TD-OCT) 310, *310*
time interval units (TIU) 186
time-resolved measurements 326–7
timers 185–6
time transfer 195–6
tissue diagnostics, applications to 327
tissue, optical properties and physical models
  absorption coefficient 207
  diffusion approximation 211–12
  ex vivo measurement 207–8, *208*
  fluence rate distributions in tissue 212
  Monte Carlo Simulation 210–11
  numerical solution of RTE 210
  radiative transport equation 210
  scattering coefficient 207
  scattering phase function 207, *208*
  in vivo measurement 208–9, *209*
tissue-selective therapy, photochemical interactions 234–5
tissue welding and soldering 335
tracking 387
  of GNSS constellations 195
training 357
transfer oscillator scheme 167
transmit/receive or monostatic systems 184
transmitters *444, 445*
  micro-sized LEDs 445–6
  organic LEDs 446
  resonant cavity LEDs 446
  white-coloured LDs 445
  white-coloured LEDs 445
transmitting optics 26, *26*
transparent tissue surgery with ultrashort-pulse lasers
  cataract surgery 237–9
  refractive surgery 237
  vitreoretinal surgery 237
transpupillary thermotherapy (TTT) 232
trapping of cold atoms
  magnetic and optical dipole traps 136
  magneto-optical trap 135–6
Tribble, J. 282
Trokel, S. L. 236
Tropea, C. 1
tumours, thermal treatment of **262**
  flexible endoscopes, laser therapy with 262–3
  gastrointestinal tract, cancers of 262
  interstitial laser photocoagulation 263–5
  lungs, cancer of 263, *263*
  thermal effects on biological tissue **261**
  urology 263
two-mode stabilization technique 5, 9–10
two-photon activation 257

ultralow-expansion (ULE) cavity 124, *124*
ultrashort-pulse lasers 237–9
ultrasound activation 258
ultra-stable optical cavities 142, *143*
Unal, M. F. 40
up-converting nanoparticles (UCNP) 257
urology 263
UV-vis-NIR photon energies 291

vaporization
  medical application 223–4
  superheated fluid expansion 223, *223*
  surface evaporation 222–3
vascular smooth muscle cells (VSMC) 273
vascular surgery, lasers in
  atheroma and angioplasty-like injury 273–4
  excimer laser angioplasty 275–6
  laser angioplasty 274–5
  PDT in arteries 276–8
velocity potential 219–20
vertical cavity surface-emitting laser (VCSELs), VLSI circuits *367*
  design modifications for high-speed P^2I^2-VCSELs 376–7
  first high-speed P^2I^2-VCSELs 377–9, *377–9*
  first implementation: The I^2-VCSEL 370–6, *371–6*
  flip-chip bonding *368*
  P^2I^2 VCSEL design 368–70, *369*
very long baseline interferometry (VLBI) systems 182
visible light communications (VLC) *447*
  digital modulation techniques 450–1
  indoor VLC channel characterizations 448–50
  multiple-input multiple-output techniques 454–6
  receivers 446–7
  transmitters 444–6
vitreoretinal surgery 237
voltage-controlled oscillator (VCO) 48
volumetric (or tomographic) (PIV/PTV) 24
vom Stein, H. D. 24
von Neumann architecture
  CMOS technologies 464
  von Neumann bottleneck 464

Wagnieres, G. A. 327
Walsh, J. T. 283
Wang, C. C. 91
'water-window' region
  imaging with lasers operating in 332–3
  imaging with soft X-rays in 331–2
wavelength-division multiplexing (WDM) 88, 428
wavelength effects 282–3
wavelength-modulated heterodyne laser interferometer 6, *6*
wavelength-selective silicon photonic devices
  AWG-based optical filters 406–7
  grating-based optical filters 410
  MRR-based optical filters 407
white-coloured LDs 445
white-coloured LEDs 445
white-light (reflectance) imaging *292,* 292–3, *293*
Wilmshurst, T. H. 48
Wriedt, T. 35
Wynne, J. 236

X-ray interferometer 7

Yeh, Y. 24

Zeeman effect 116
Zenisek, D. 329
Zhang, J. 326

Taylor & Francis Group
an **informa** business

Taylor & Francis eBooks

www.taylorfrancis.com

A single destination for eBooks from Taylor & Francis with increased functionality and an improved user experience to meet the needs of our customers.

90,000+ eBooks of award-winning academic content in Humanities, Social Science, Science, Technology, Engineering, and Medical written by a global network of editors and authors.

TAYLOR & FRANCIS EBOOKS OFFERS:

A streamlined experience for our library customers

A single point of discovery for all of our eBook content

Improved search and discovery of content at both book and chapter level

REQUEST A FREE TRIAL

support@taylorfrancis.com

Routledge
Taylor & Francis Group

CRC Press
Taylor & Francis Group